



HAL
open science

QWIN : système automatisé de prédiction, trading et placement de paris sportifs

Julien Teisseire

► **To cite this version:**

Julien Teisseire. QWIN : système automatisé de prédiction, trading et placement de paris sportifs. Informatique et théorie des jeux [cs.GT]. 2017. dumas-03356129

HAL Id: dumas-03356129

<https://dumas.ccsd.cnrs.fr/dumas-03356129>

Submitted on 27 Sep 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

CONSERVATOIRE NATIONAL DES ARTS ET METIERS
PARIS

QWIN
Systeme automatisé de prédiction,
trading et placement de paris sportifs

Mémoire de validation du diplôme d'ingénieur DPE
Spécialité Informatique

Julien Teisseire

04/09/2017

Ecole

CONSERVATOIRE NATIONAL DES ARTS ET METIERS
PARIS

<http://ecole-ingenieur.cnam.fr/>



Composition du Jury

Président	Madame le Professeur Elisabeth METAIS	Représentant l'administrateur général du CNAM
Enseignants	Monsieur Eric-Nazir ABOU CHAKRA Monsieur Cédric DU MOUZA	Professeur associé CNAM Maître de conférence CNAM
Ingénieurs	Monsieur Lionel ICARD Monsieur Christophe GODON	Ingénieur Ingénieur

Table des matières

1	Préface	5
2	Remerciements	6
3	A propos de l’auteur	7
4	Besoin	8
4.1	Le problème	8
4.2	Le cahier des charges	10
5	Solution	11
5.1	Phase exploratoire	11
5.1.1	Proof of concept	11
5.1.2	Revente des pronostics	15
5.2	Etudes techniques et fonctionnelles	18
5.2.1	Collecte de résultats	18
5.2.2	Plate-forme de betting	18
5.3	Big picture	20
5.4	Architecture	21
6	Environnement	22
6.1	Macbook workstation	22
6.2	Git	22
6.3	VM Linux Rackspace	23
6.4	VPN	24
6.5	PHP et C++	25
6.6	MySQL Cluster	25
6.7	Apache et HHVM	26
7	Spécifications	28
7.1	Stratégie de pari	28
7.2	Big Data	29
7.2.1	Collecte des données	29
7.2.2	Traitement des données	32
7.3	Simulation	33

7.3.1	Prédictions avant-match	33
7.3.2	Prédictions inplay	34
7.3.3	Simulateur de résultats	34
7.4	Prédiction	35
7.4.1	Prédictions avant-match	35
7.4.2	Prédictions inplay	35
7.5	Mode de pari	36
7.5.1	Les différents types de marché	36
7.5.2	Pari de type BACK	37
7.5.3	Pari de type LAY	37
7.5.4	Notion de Bucket et fonctionnement de l'Exchange	38
8	Conception et développement	41
8.1	Présentation de l'API Betfair	41
8.2	API Qwin - Webservice interne	43
8.3	QwinBoard - Tableau de bord	45
8.4	BFSstats - Module de collecte des événements	47
8.5	QwinBet - Module de prédiction et placement des paris avant match	47
8.6	IPControl - Module de collecte et de placement des paris "inplay"	47
8.7	QwinRank - Module de ranking des players	48
8.8	Popularity - Module de calcul de popularité	48
8.9	Rewind - Module de reporting et simulation	48
8.10	Module de gestion de librairies internes et de la base de données	50
9	Tests et Optimisations	53
9.1	Tests unitaires	53
9.2	Tests de non-regression et logs	54
10	Mise en production	55
10.1	Pré-production	55
10.2	Real betting	56
11	Livraison et sauvegarde	57
11.1	Livraison	57
11.2	Sauvegarde	57
12	Recherche d'algorithmes	58
12.1	Phase exploratoire	58
12.2	Essais en simulateur	58
12.3	Passage en production	59
13	Bilan financier	60
13.1	Budget initial	60
13.2	Gestion de la trésorerie	61

14 Conclusion	62
14.1 Synthèse financière	62
14.2 Synthèse technique et organisationnelle	63
14.3 Synthèse humaine	64
15 Annexes	65
15.1 Grandes phases de l'approche fonctionnelle du projet	66
15.2 Exemples de placement de pari sur l'API Betfair	67
15.3 Extrait des logs du système Qwin	68
15.4 Trading Inplay, apprentissage et modèles prédictifs	70
15.5 Rewind, exemple de résultats en sortie du simulateur	72
15.6 Memento des commandes de configuration OpenVPN	77

Chapitre 1

Préface

Ce mémoire constitue une synthèse des étapes réalisées dans le cadre d'un projet professionnel de trading "haute-fréquences" dans le domaine des paris sportifs.

Entre mes différentes missions et emplois, j'ai choisi de démarrer un projet autour d'une activité de trading, sans client direct, à fort potentiel et forte complexité technologique (big data, performance, prédiction/Business Intelligence).

C'est sur la base d'une équipe polyvalente de 2 personnes, mon associé et moi-même que l'aventure a démarré.

Ce challenge a permis de comprendre les mécanismes internes des places de marché dans le monde des paris sportifs.

Je vais présenter ici comment le projet a vu le jour, comment ce dernier a été réalisé (toutes les phases de gestion d'un projet d'ingénieur), et quelles ont été les conclusions techniques et financières de ce dernier.

Habitué à mener des projets pour mes clients ou au sein même de mon entreprise, j'ai consacré 2 années de 2013 à 2015 à la conception et au développement de l'intégralité du système Qwin et mis à disposition l'ensemble des moyens logistiques nécessaires. Ce projet a permis l'accomplissement d'un véritable défi technique en plus d'être une véritable découverte d'un univers que je ne connaissais pas, celui des places de marchés sportifs.

Chapitre 2

Remerciements

Dans le cadre de ce mémoire, en vue de l'obtention du titre d'ingénieur diplômé par l'état, je souhaitais adresser mes remerciements à plusieurs personnes.

Le projet que je décris ici a été un projet important à bien des égards.

J'ai mené énormément de projets de développement durant ma carrière pour différents besoins et en utilisant différentes technologies. Le travail entre ingénieurs depuis des années m'a apporté les connaissances nécessaires à l'accomplissement de ce projet complexe et atypique. Complexe car sur des bases de connaissances nouvelles et atypique car relevant d'un désir d'entreprendre, sur fonds propres, dans un domaine particulier souvent associé au jeu et parfois à la dépendance pour certains.

C'est dans ce contexte qu'il a fallu expliquer et surtout démontrer qu'il ne s'agissait nullement d'un jeu mais réellement d'un projet professionnel concret, pour soi-même et pour les autres.

Pour cette compréhension et ce soutien je tiens à remercier ma compagne Najette Boutenzourht qui a su être patiente et toujours compréhensive dans l'investissement en temps et en moyen que représentait ce projet.

Je tiens à remercier mon partenaire Frédéric Leroy sur le projet avec qui il a été possible comme c'est rarement le cas de passer par les moments les plus difficiles comme les plus heureux, sans jamais entacher notre amitié.

Je tiens également à remercier tous mes proches, amis, famille pour leurs questions et l'intérêt apporté à ce projet.

Un grand merci à mon ancienne responsable Eurielle Trombert durant mon activité au sein de l'entreprise Orange qui a toujours compris mes choix et m'a encouragé, à contre-cœur certes, dans ma décision de quitter mon poste pour réaliser ce projet.

Et enfin un grand merci au jury de cette validation des acquis pour le temps consacré, les échanges que nous avons eu durant l'entretien d'évaluation et leur conseil de choisir ce projet particulièrement dans le cadre de mon mémoire, plus que d'autres sujets tout aussi passionnant mais certainement plus conventionnels.

Chapitre 3

A propos de l'auteur

A la suite d'un cursus d'études m'amenant au niveau Bac+4 (Bac S option Technologies Industrielles / BTS Informatique Industrielle / DU Ingénierie des Réseaux Informatiques) et après l'obtention de différentes certifications professionnelles (Nortel Networks TM, Sun Microsystems TM, Microsoft TM, Scrum.org TM), j'ai directement dû rentrer dans la vie active pour différentes raisons. Mon désir d'obtenir le titre d'ingénieur et ma motivation ont commencé dès mes débuts professionnels, mais ma situation personnelle m'a contraint à avancer et à construire progressivement ma carrière, en gardant à l'esprit ce sentiment d'inachevé.

Travaillant depuis toujours comme ingénieur, leader technique ou encore chef de projet, et au sein d'équipes d'ingénieurs, le besoin de reconnaissance relève d'une réelle aspiration personnelle, plus que d'une nécessité professionnelle.

Mon expérience s'allongeant, depuis l'année 2000, c'est à deux reprises que j'ai cherché à obtenir ce chaînon manquant dans mes études. En 2002, avec un travail du soir auprès du CNAM (Intégration Client/Serveur) puis en 2010 toujours auprès du CNAM. Il a été difficile d'aller au bout de la démarche. Toujours pour des raisons personnelles, le contexte ne réunissait pas les conditions adéquates pour ce type de démarche.

Après la création de deux sociétés et des emplois dans différentes entreprises comme Quescom, Amadeus, Orange, la Caisse d'épargne ou encore Monaco Télématique, ma détermination pour réaliser cette partie manquante de mon cursus est intacte.

Actuellement en poste chez Thales Services, au titre d'architecte logiciel, je souhaite sereinement accomplir cette importante étape de ma vie.

Après la constitution de mon dossier et l'entretien réussi devant le jury, je suis honoré de vous présenter ce mémoire et d'en assurer la soutenance en vue, je l'espère, de l'obtention du titre d'ingénieur DPE, spécialité Informatique.

Chapitre 4

Besoin

4.1 Le problème

La caractéristique principale du projet étant d'être son propre donneur d'ordre, le problème auquel il a fallu répondre était le suivant : Être en mesure de générer des revenus au-travers d'une activité de placement de paris sportifs en s'appuyant sur notre capacité à résoudre l'équation suivante :

("Prédictions" x "Gains") - ("Prédictions" x "Pertes") = "Profit" - "Frais" = "Revenus".

Sur la base de ce problème, le besoin était d'assurer les meilleures prédictions possibles, et évidemment, au niveau des gains et des pertes, de travailler autour des meilleures côtes.

Dès le départ il était prévu de ne pas devenir trader, mais de développer un ensemble d'automates pour assurer les calculs et la gestion complète de l'activité de trading de façon automatisée.

Dans un contexte très fortement réglementé, les composantes suivantes ont été à prendre en compte :

- Législation : Probablement l'aspect le plus contraignant du projet. Les pays ont tous une réglementation différente et fortement restrictive. L'objectif affiché est la protection des parieurs, mais en réalité tout est défini pour garantir le passage à l'imposition. Les états cherchent à contrôler les échanges financiers autour des paris sportifs et assurer des revenus sur les recettes. C'est pour cela que la mise en place d'une solution légale a rajouté en complexité technique.

Le choix a donc été pris de travailler avec une plate-forme de trading en Angleterre. Cela a été possible grâce à la mise en place d'un VPN et l'utilisation du cloud Rackspace hébergé outre-manche.

- Concurrence : Il existe énormément de sites internet ou de conseillers en paris sportifs. Nous avons travaillé au début sur cet axe business, mais compte-tenu des résultats, c'est la réalisation de pronostics pour notre propre besoin qui a été privilégiée. A partir de ce principe, la concurrence faisait partie intégrante de notre business, tout comme en bourse ou chaque trader "joue" contre les autres. La concurrence naturelle était en résumé l'ensemble des utilisateurs de la plate-forme Betfair.

- Faisabilité technique : A l'heure du Bigdata, des technologies NoSQL et des capacités de calcul distribué qu'offrent les frameworks tels que Spark, l'inquiétude pour ce projet ne se portait pas sur les outils qu'il fallait mettre en place, mais plus sur notre capacité à définir les critères principaux et leurs valeurs, permettant de détecter les événements sur lesquels parier et à quel moment le faire. Les compétences au démarrage du projet étaient orientées :
 - MySQL Cluster pour le SGBD (utilisé par le passé chez Orange France Télécom)
 - GlusterFS pour le système de fichier distribué (expérimentation et montée en compétences dans notre laboratoire de travail)
 - HDFS/Hadoop pour l'exécution des tâches de traitement (utilisé pour le filtrage et la consolidation des historiques de résultats et des côtes). Utilisé à titre expérimental mais pas en production
 - Php, JQuery et C++ en ce qui concerne les langages de programmation (langages les mieux maîtrisés qui ont permis la réalisation du projet)

C'est sur cette base technique que j'ai pu commencé les premières expérimentations et ensuite travaillé sur la définition fonctionnelle détaillée des différents modules qui constitue la solution finale.

- Budget/Délai Il est impossible de parler du projet d'un point de vue technique sans évoquer également les aspects organisationnels. Le projet a été établi avec une vision moyen terme, un financement sur 2 ans et une trésorerie d'environ 100 000 euros. La période sans financement annexe (sans mission de consulting parallèle) a duré 15 mois. Ce budget et ce délai ont permis de disposer de suffisamment de temps et de moyen pour assurer une infrastructure technique efficiente, une rémunération constante et surtout une capacité à risquer des sommes importantes (jusqu'à 20 000 euros par journée de paris) sur des phases de trading (je reviendrai plus loin dans les détails sur les stratégies de placement de paris).

4.2 Le cahier des charges

A partir de ce qui a été exposé plus haut, le contexte a clairement été la mise en place d'un projet regroupant les exigences suivantes :

- Tenir la réserve budgétaire
- Disposer des données complètes de l'ensemble des évènements sportifs à l'international
- Disposer des données complètes de l'ensemble des côtes pour chacun des participants de l'ensemble des évènements sportifs
- Disposer d'une capacité de rétention des données pour une fenêtre de 10 ans
- Disposer d'une plate-forme d'analyse des données et de simulation des paris en fonction d'un ensemble d'algorithmes à concevoir
- Concevoir une série d'algorithmes pour étudier les différents profils de trading envisageables
- Disposer d'un robot d'analyse de flux temps-réel pour la génération de pronostics et la détection d'opportunités
- Disposer d'un robot de placement des paris sur la base d'une stratégie financière et des prédictions calculées
- Disposer d'un tableau de bord de l'ensemble des opérations
- Disposer d'un mécanisme de sauvegarde et d'archivage de la solution et des données
- Disposer d'un module de gestion du risque et de création d'alertes

Sur la bases de ces contraintes et exigences définissant notre cahier des charges initial, le projet a pu démarrer par les phases de validation et de Proof of Concept. En mode itératif, chaque étape a permis de passer à la suivante, en prenant soin systématiquement de ne pas entrer dans un effet tunnel. Ainsi, il a fallu fixer sprint après sprint, des objectifs simples et rapides, sur lesquels la gestion en amélioration continue peut s'appliquer.

Dans ce mode de fonctionnement, il existe forcément un sur-coût (overhead) lié parfois à des nécessités de reprises de code (refactoring). Mais par rapport à la réduction des risques que cette dernière induit, elle reste la méthode la plus adaptée à ce type de projet d'exploration et de développement.

Chapitre 5

Solution

5.1 Phase exploratoire

5.1.1 Proof of concept

Crawl des sites de statistiques

La première étape fût la mise en place d'un Proof Of Concept.

En utilisant différents sites de résultats sportifs et de classement, une première version de la base de données a pu être constituée.

Les sites utilisés, entre autres, étaient :

- <http://fussball.wettpoint.com/en/>
- <http://www.betexplorer.com>

Un premier développement a permis l'exploration de la structure de ces sites (crawl + parsing du DOM html) et la collecte des données en masse semaine après semaine.

Dans cette étape de test, seulement le football a été étudié.

Exemple de l'utilisation de cURL en PHP pour la connexion aux sites de statistiques sportives :

```
// Collect next scheduled matches
function connectWettPoint($pDay, $pMonth, $pYear, $pLeague)
{
    //open connection
    $ch = curl_init();

    //set the url, number of POST vars, POST data
    // curl_setopt($ch,CURLOPT_URL, "http://213.175.197.228/en/schedules/");
    curl_setopt($ch,CURLOPT_URL, "http://fussball.wettpoint.com/en/schedules/");
    curl_setopt($ch,CURLOPT_POST, 1);
    $lDataString =
    "tag=".$pDay."&monat=".$pMonth."&jahr=".$pYear."&liga=".$pLeague."&lsort=0&B1=Show%20Matches";
    curl_setopt($ch,CURLOPT_HTTPHEADER, array('Content-Length: '.strlen($lDataString)));
    curl_setopt($ch,CURLOPT_POSTFIELDS, $lDataString);
}
```

```
curl_setopt($ch,CURLOPT_HEADER, 0);  
curl_setopt($ch,CURLOPT_RETURNTRANSFER, 1);  
...
```

Pour éviter d'être *blacklisté* et donc interdit de contacter les sites de résultats et classements, il a fallu élaborer un programme plus intelligent de gestion de connexions réseaux et des chargements HTML par l'intermédiaire de Proxies pour contourner les contrôles côté serveurs (car le volume d'informations collectées dépassé le cadre de la simple visite d'un utilisateur classique pour ces sites).

L'utilitaire ainsi développé était en mesure de temporiser entre les requêtes, de traiter les résultats via un parsing du DOM (conversion en objet), de normaliser les données avant de les insérer en base.

Traitement CSV

Les données extraites durant la phase d'aspiration des sites d'informations sportives étaient en synthèse :

- Le classement de chaque équipe
- Le nombre de point de chaque équipe
- Le nombre de journée dans le championnat
- La journée en cours
- Les informations disponibles annexes : match à l'extérieur, match à domicile
- Les côtes de chaque bookmaker pour 1N2 (en BACK et en LAY, explication des termes plus loin dans le mémoire)

Les données au format CSV ont permis une injection facilitée dans le cluster MySQL.

Le volume de données et leur sécurisation étaient assurés par 4 machines virtuelles Linux Fedora partageant un volume GlusterFS.

Le travail a été effectué sur l'ensemble des équipes de ligue 1 sur les principaux pays européens (Angleterre, Allemagne, Italie, France, Espagne, Portugal).

Prédiction basiques

Les toutes premières analyses et prédictions étaient réalisées manuellement puis grâce à un script . C'est à l'aide d'un fichier Excel que les paris étaient suivis et comptabilisés. Cette phase du projet a servi de Proof Of Concept fonctionnel, la possibilité de travailler autour de statistiques pour la prédiction de résultats sportifs.

Exemple de fichier Excel utilisé pour le suivi des paris :

Total Bet for empowered = 2/301

LD:3/40 Date:16/05/13, 21:15 Fylkir(pts:1,rk:8,diff:-1) Stjarnan(pts:3,rk:7,diff:+0)	Iceland - Pepsi Deildin Nul match	PMU 500	PERDU 😞
LD:3/40 Date:20/05/13, 21:15 Keflavík(pts:0,rk:11,diff:-3) Fylkir(pts:1,rk:8,diff:-1)	Iceland - Pepsi Deildin Nul match		

Total Bet for golum = 19/301

LD:32/36 Date:16/05/13, 19:45 Basel(pts:63,rk:1,diff:+27) Lausanne(pts:27,rk:9,diff:-20)	Switzerland - Super League Basel will win	Cote trop basse Gagné	Combo du jour pas interessant
LD:3/40 Date:16/05/13, 20:00 KR(pts:6,rk:4,diff:+3) Thor(pts:0,rk:12,diff:-6)	Iceland - Pepsi Deildin KR will win	Cote trop basse Gagné	Combo du jour pas interessant
LD:37/38 Date:17/05/13, 20:30 Nantes(pts:65,rk:2,diff:+24) Sedan(pts:30,rk:19,diff:-16)	France - Ligue 2 Nantes will win	Pari non trouvé Gagné	
LD:37/38 Date:17/05/13, 20:30 Monaco(pts:70,rk:1,diff:+29) Le Mans(pts:37,rk:17,diff:-23)	France - Ligue 2 Monaco will win	Gagné	PMU 1000 BETCLIC 1000
LD:37/38 Date:18/05/13, 21:00 Paris SG(pts:77,rk:1,diff:+42) Brest(pts:29,rk:20,diff:-27)	France - 1. Ligue Paris SG will win	Gagné	
LD:34/34 Date:18/05/13, 15:30 Borussia Dortmund(pts:66,rk:2,diff:+40) 1899 Hoffenheim(pts:28,rk:17,diff:-26)	Germany - 1. Bundesliga Borussia Dortmund will win	Perdu	BETCLIC 2000
LD:34/34 Date:18/05/13, 15:30 Borussia Mönchengladbach(pts:47,rk:8,diff:-3) Bayern München(pts:88,rk:1,diff:+79)	Germany - 1. Bundesliga Bayern München will win	Gagné	
LD:38/38 Date:18/05/13, 20:45 Sampdoria(pts:40,rk:15,diff:-9) Juventus(pts:87,rk:1,diff:+48)	Italy - Serie A Juventus will win	Perdu	BETCLIC 1000 UNIBET 1000

LD:34/36 Date:18/05/13, 18:30 Red Bull Salzburg(pts:70,rk:2,diff:+47) FC Wacker Innsbruck(pts:30,rk:10,diff:-32)	Austria - Bundesliga Red Bull Salzburg will win	Gagné	BETCLIC 1000
LD:34/36 Date:18/05/13, 18:30 FK Austria Wien(pts:76,rk:1,diff:+50) SC Wiener Neustadt(pts:30,rk:9,diff:-31)	Austria - Bundesliga FK Austria Wien will win	Gagné	
LD:38/38 Date:19/05/13, 17:00 Liverpool(pts:58,rk:7,diff:+27) Queens PR(pts:25,rk:20,diff:-29)	England - Premier League Liverpool will win	gagné	
LD:38/38 Date:19/05/13, 17:00 Manchester C(pts:78,rk:2,diff:+33) Norwich(pts:41,rk:12,diff:-18)	England - Premier League Manchester C will win	Perdu	BETCLIC 1000 PARIONSWEB 1000
LD:38/38 Date:19/05/13, 17:00 West Bromwich(pts:48,rk:8,diff:-4) Manchester U(pts:88,rk:1,diff:+43)	England - Premier League Manchester U will win	Perdu	
LD:38/38 Date:19/05/13, 20:45 Pescara(pts:22,rk:20,diff:-53) Fiorentina(pts:67,rk:4,diff:+24)	Italy - Serie A Fiorentina will win	Gagné	BETCLIC 1000
LD:36/38 Date:19/05/13, 21:00 FC Barcelona(pts:91,rk:1,diff:+69) Valladolid(pts:43,rk:12,diff:-4)	Spain - Primera Division FC Barcelona will win	Gagné	
LD:30/30 Date:19/05/13, 19:30 Benfica(pts:74,rk:2,diff:+55) Moreirense(pts:24,rk:15,diff:-19)	Portugal - Liga Zon Sagres Benfica will win	Gagné	
LD:32/34 Date:19/05/13, 19:30 Steaua Bucuresti(pts:75,rk:1,diff:+44) Viitorul Constanta(pts:30,rk:14,diff:-15)	Romania - Divizia A Steaua Bucuresti will win	Non trouvé Perdu	
LD:29/30 Date:19/05/13, 12:30 Metalurh Zaporizhya(pts:10,rk:16,diff:-49) FC Shakhtar Donetsk(pts:75,rk:1,diff:+60)	Ukraine - Premiere League FC Shakhtar Donetsk will win	Perdu (nul)	
LD:8/40 Date:20/05/13, 18:30 FC Infonet(pts:10,rk:7,diff:-10) Levadia Tallinn(pts:28,rk:2,diff:+13)	Estonia - Meistriliiga Levadia Tallinn will win	Non trouvé	

6 perdus sur 18

Algo = 66% ...

JU Capital = 25k€ :: 2200 € par combo/nul et 180 € par single
TOTAL PRUDENCE + GAVAGE BWIN CETTE SEMAINE → Mise = 10500 €

FRED Capital = 12k€ :: 1100 € par combo/nul et 90 € par single
Mise = 10500 k€

5.1.2 Revente des pronostics

Prediday - Sites de revente des pronostics

Afin de consolider la trésorerie et valoriser les pronostics générés, il a été décidé de concevoir un site internet de pronostics.

Le site a été réalisé en utilisant les technologies Fullstack classiques telles que : PHP pour le backend, jQuery et HTML5 pour le frontend et MySQL pour la base de données.

Une librairie Paypal a été intégrée afin de garantir des paiements sécurisés sur la plate-forme.

La partie publique du site proposait des pronostics du jour avec des algorithmes peu optimisés. Cela avait pour but d'inciter l'utilisateur de Prediday à payer afin d'obtenir les pronostics basés sur des algorithmes plus performants.

Les paris potentiellement achetables sont symbolisés par un verrou. Seuls le pays et la côte sont indiqués afin de ne pas dévoiler trop d'informations à l'utilisateur qui pourrait deviner le match sur lequel parier.

The screenshot shows the Prediday website interface. At the top, there's a header with the logo and the text 'prediday.fr'. Below the header, there's a user profile section with 'sample@email.tst', 'Credits : 0', and buttons for 'Obtenir des crédits' and 'Logout'. The main content area is titled 'Le meilleur tipster football' and displays three line charts showing success rates for 'public' (53%), 'privé' (67%), and 'gold' (54%) categories. Below the charts, there's a section for 'Matches du 18/03/2014' with a note '(mise sur les équipes en vert)'. There are three betting grids: 'Grille 'Gold'', 'Grille 'Privée'', and 'Grille 'Publique''. The 'Grille 'Gold'' and 'Grille 'Privée'' grids show two matches with a 'Verrou' (lock) icon and a message 'Voir ce pronostic pour 2 crédits !'. The 'Grille 'Publique'' grid shows two matches: 'Real Madrid vs Schalke' (1.23) and 'Crawley Town vs Wolves' (1.79). At the bottom, there's a disclaimer: 'Parier comporte des risques : Endettement, isolement, dépendance. Pour être aidé, appelez le 09-74-75-13-13 (appel non surtaxé). Référez vous aux conditions générales des sites de paris en ligne #45365 à propos de prediday'.

FIGURE 5.1 – Prediday, Portail de vente de pronostics

L'achat de crédits

Lorsque l'utilisateur du site le souhaitait, il pouvait acheter des crédits via paypal (à terme d'autres moyens de paiement). Ces crédits lui permettaient de découvrir les paris « gold » et « private ».



FIGURE 5.2 – Prediday, Achats de pronostic via Paypal

Les parties « gold » et « private »

Les parties gold et private représentent les pronostics basés sur des algorithmes performants. Le prix diffère entre gold et private. Le pourcentage de prédiction en gold étant supérieur au pourcentage de prédiction en private. Le prix d'une prédiction gold était alors supérieur au prix d'une prédiction private (la monnaie est en « crédits »).

Lorsque l'utilisateur possède des crédits, un simple clic sur le match dévoile le pronostic. (Exemple match Chelsea – Galatasaray sur l'image ci-dessous).

L'utilisateur pouvait alors parier le match avec son bookmaker.



FIGURE 5.3 – Prediday, Grilles disponibles en fonction des probabilités des pronostics

Expérimentation grande échelle

Le seul objectif de ce site internet était d’augmenter la trésorerie, tout en se laissant le temps de travailler sur les algorithmes.

Son fonctionnement a été convainquant car il a confirmé notre capacité à produire des pronostics fiables. La difficulté est venue dans le cadre des paris haute-fréquence, et en mode inplay.

C’est à cet instant que la limite des paris avant-match est apparue. Il a fallu se concentrer sur la collecte d’informations et le traitement à grande échelle, expliqués plus loin.

5.2 Etudes techniques et fonctionnelles

Comme pour chaque projet, une partie étude technique est nécessaire afin de faire les choix les plus pertinents dans le but d'assurer le succès et la longévité de ce dernier.

5.2.1 Collecte de résultats

Statistiques par sport post-match

Pour réaliser la collecte des données, plusieurs possibilités ont été étudiées. La première présentée dans cette section réside en la récupération des données des sites internet dédiés au football et quelques autres sports (ex : Wettpoint).

La faible quantité de données statistiques disponibles a limité les travaux de prédictions à 2 sports en particulier, le football et le tennis.

J'ai alors procédé au développement des premiers outils de parsing HTML, sur les sites suivants :

- Les données du site Wettpoint
- Les données du site betexplorer.com
- Les données récupérées des archives CSV de Betfair

Les derniers algorithmes de production n'utilisaient finalement plus que les données financières ainsi que les côtes du marché. Les côtes intégrant en définitive déjà tous les éléments statistiques et probabilistes, il ne s'agissait plus que de déterminer le meilleur instant pour parier.

Données temps-réelles

Avec le développement de notre connecteur Betfair, SOAP/XML puis REST/Json, il a été possible d'obtenir l'ensemble des informations disponibles et ainsi de les stocker en parallèle des traitements "temps-réel" pour des analyses postérieures. Le terme "temps-réel" utilisé ici signifie que sur une exigence de temps donnée et à intervalle régulier, le système lançait ses propres évènements pour la capture des marchés et des évolutions de côtes.

5.2.2 Plate-forme de betting

Bookmakers français

La phase d'étude a aussi donné lieu à une analyse précise des possibilités proposées pour le placement de paris en France. La liste des bookmakers français est longue et pour être certain dans la phase de recherche de la meilleure plate-forme de disposer des meilleures côtes, il a fallu créer des comptes sur l'ensemble des sites de bookmakers. Chaque bookmaker avait ses spécificités, et des côtés proches mais basées sur leurs propres algorithmes et donc légèrement différentes. Liste des bookmakers qualifiés : Betclac, Bwin, PMU, Netbet, Unibet.

La conclusion de notre phase de tests est que les bookmakers français ont pour vocation de gagner de l'argent sur l'issue des matchs, et pas simplement sur le service proposé.

Il ne s'agit en rien de place de marché ou les traders s'affrontent pour remporter la mise, mais bien d'une logique non-équitable de gestion des revenus des bookmakers. Les côtes sont donc calculées chez ces derniers pour, en fonction des probabilités, garantir des pertes minimales et des profits importants.

C'est une opération simple pour un bookmaker qui va chercher à inciter les parieurs à investir sur des événements en maintenant la balance financière à leur avantage (maîtrise des côtes).

Bookmakers étrangers

A l'inverse, dans le cadre de l'étude de viabilité du projet, les recherches ont conduit à Betfair et Betdaq.

Ces deux bookmakers anglais se comportent comme de véritable place de marché, où il n'y a certes pas de limite dans les pertes, mais pas non plus de limite dans les gains.

Ce fonctionnement plus proche de la bourse que du jeu en ligne permet aux bookmakers d'en vivre car ils se comportent également comme Exchange, c'est à dire comme de véritable place de marchés.

En chargeant les profits par une commission, l'Exchange est garanti de gagner de l'argent sans que cela ne soit à l'encontre des traders.

C'est grâce à ce fonctionnement que le système devient équitable, et au-travers d'une API d'accès aux données et de placement des ordres il devient pleinement envisageable de construire un business autour de cette activité.

Comme détaillé dans la section législation, ce genre de place de marchés sportifs est interdite en France.

L'explication est simple, étant donné que l'état n'est pas en capacité de capter des revenus sur cette activité, il préfère l'interdire, au prétexte de défendre les intérêts des parieurs en leur évitant de mettre leur épargne à risque.

C'est ainsi que les bookmakers que l'on connaît ont pu prospérer.

Petite anecdote, en pariant au démarrage du projet sur les sites des bookmakers français, j'ai été purement et simplement limité puis interdit de pari suite à d'importants gains générés. Les sites français nous protègent contre l'endettement mais contre l'enrichissement personnel également. Mieux vaut ne pas être trop performant sur ces plate-formes ...

L'étude a donc permis de conclure sur la nécessité d'utiliser une réelle plate-forme de type place de marchés (Exchange), uniquement disponible aux résidents anglais et à quelques pays encore, pour peu de temps, ouverts à cette pratique.

Les API Betfair et Betdaq ont été analysées et la conclusion a été que Betfair propose davantage de marchés (sports différents) et d'évènements (type de paris par sport). Le nombre d'utilisateurs étant également plus conséquent, les volumes monétaires en jeux sur chacun des événements permettaient de quasiment systématiquement prévoir des ordres.

Information sur l'API de Befair : <https://www.programmableweb.com/api/betfair>

5.3 Big picture

La Big Picture est l'élément clé d'un projet. Elle reprend l'ensemble des modules et des flux de traitement. La Big Picture permet de disposer de la vision globale de l'architecture. Elle permet également de partager avec toutes parties de l'équipe projet, la même vision et le même objectif final.

Sur cette base, il devient possible de découper les travaux et de procéder à la phase de conception logicielle en répartissant convenablement l'ensemble des classes à développer.

La solution qui a été retenue sur la dernière version de production consiste en l'utilisation des données directement collectées par le robot de placement de paris. Les données statistiques existantes permettaient une pondération et une vérification du comportement des traders.

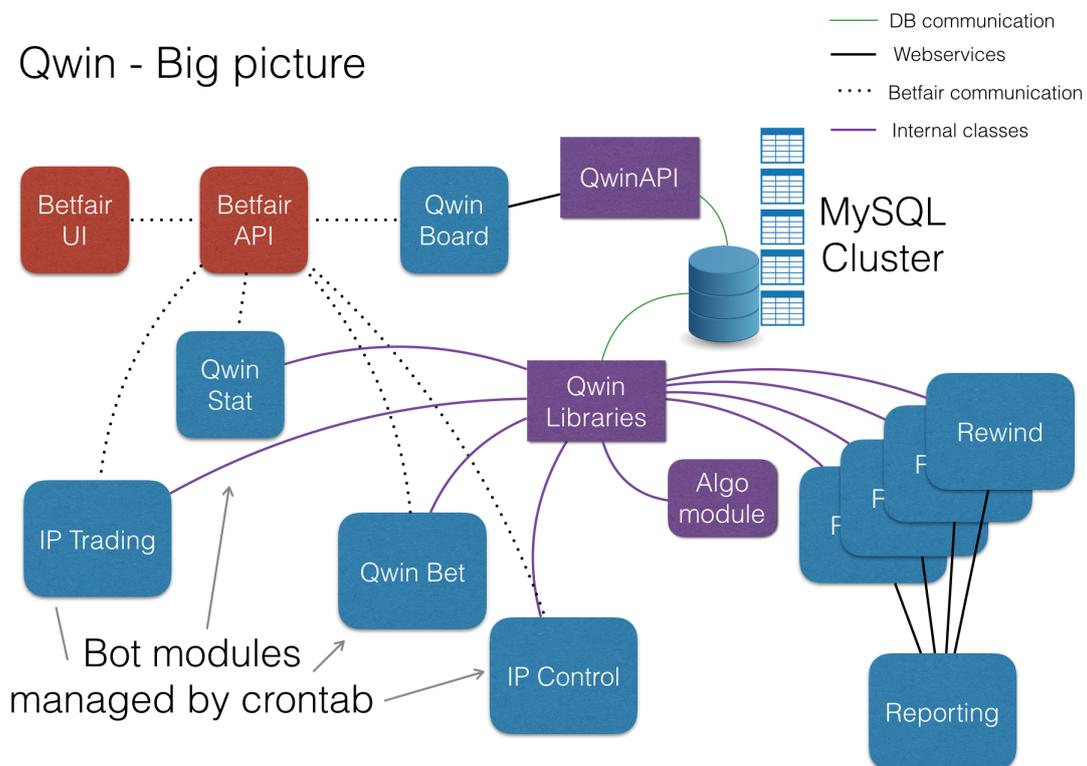


FIGURE 5.4 – Qwin, Big Picture

5.4 Architecture

Architecture technique

Les serveurs étaient hébergés chez Rackspace en Angleterre. Plusieurs entrées dans la Crontab assuraient l'exécution des jobs et un mécanisme de compression et transfert de la base de données en France était en place pour assurer la sauvegarde des données d'une part, et la mise à disposition de données fraîches sur notre machine de développement d'autre part.

Le problème de la volumétrie s'est vite posé et le cluster MySQL hébergé chez Rackspace est rapidement devenu indispensable pour garantir les performances "temps-réelles".

Voici le découpage des composants utilisés pour réaliser l'infrastructure technique du projet :

- 2 machines de développement de type MacBook Pro Core i7 / 16Go RAM / 512Go SSD
- 1 serveur MacPro de sauvegarde locale des données et du code source (serveur Git)
- 1 VPN entre notre réseau en France et le Portugal pour les accès à Betfair
- 2 serveurs de type Frontaux Fedora pour l'exécution des programmes de production et pré-production
- 1 cluster MySQL de Production pour le stockage des données

Architecture logicielle

L'architecture logicielle disposait de 2 approches.

L'architecture de type 3 tiers pour l'outil d'administration/supervision QWinBoard : FrontEnd - BackEnd API - Database.

L'architecture de type batch processing pour les automates : Crontab - Batch serveur - BackEnd API - Database.

Chapitre 6

Environnement

6.1 Macbook workstation



FIGURE 6.1 – OSX, le logo

Le choix a été fait de travailler sous Mac compte-tenu de la directe compatibilité avec les outils Linux habituellement utilisés. L'absence d'utilisation d'une machine linux virtuelle type Fedora sous Virtual Box est un confort qu'il a été décidé d'adopter. De plus le support natif de Apache a facilité les phases de développement et la rapidité de mise en oeuvre du POC. L'éditeur choisi pour le projet est Sublime Text.

6.2 Git



FIGURE 6.2 – Git, le logo

L'utilisation de Git comme gestionnaire du code source a été un choix évident. Historiquement habitué à utiliser CVS puis Bazaar, j'avais également découvert Git précédemment. Le fonctionnement décentralisé facilite l'administration car il ne dépend pas directement d'un serveur. En paramétrant des dépôts distants, il permet la sécurisation du code source et la synchronisation des environnements de développement.

6.3 VM Linux Rackspace



FIGURE 6.3 – Rackspace, le logo

Concernant Rackspace, cela a été un choix par défaut. En effet, avec la contrainte forte de disposer de serveurs au plus près de notre Exchange Betfair, la seule solution fiable pour une gestion à distance de notre infrastructure était Rackspace.

Amazon (AWS) ne disposant pas de centre d'hébergement dans cette région, et malgré les coûts plus élevés pratiquaient par Rackspace, c'est sur cet hébergeur que le choix s'est porté, dans le datacenter de Londres.



FIGURE 6.4 – Red Hat, le logo

L'utilisation de Docker a été envisagé, mais ne représentait pas d'intérêt direct pour le projet. Malgré la création d'une image Docker comprenant le WebService QWin (à titre expérimental), il a été décidé de conserver l'infrastructure en place composée de 2 frontaux Apache sur les VM Rackspace (Linux RedHat).

Une VM complémentaire, toujours sous RedHat, servait de Bastion d'entrée dans notre infrastructure et 3 VM étaient dédiées aux instances MySQL Cluster. La fonction d'auto-sharding était configurée pour la distribution des données entre les nodes MySQL. Cela signifie que les nodes servaient à améliorer la performance du cluster. La sauvegarde était assurée comme détaillé plus bas, sur une machine locale en France.

6.4 VPN



FIGURE 6.5 – HideMyAss, le logo

L'utilisation de HMA a permis de maintenir les connexions depuis la France vers le Portugal en étant considéré comme résident portugais pour l'utilisation d'un compte autorisé sur la plateforme Betfair.

Ce montage a été la plus importante contrainte (disposer d'un compte Betfair depuis une identité portugaise et un compte bancaire portugais, accéder à notre compte Betfair à partir d'une connexion Internet depuis le Portugal).

Betfair était légal au Portugal et je ne connaissais pas de personne de confiance en Angleterre. Pour faciliter les développements en local et se passer du logiciel HMA, un serveur OpenVPN a ensuite été configuré sur un VM Rackspace afin de faire rentrer nos machines en France directement dans le réseau de l'hébergeur londonien.

Voir annexe "Memento des commandes de configuration OpenVPN".

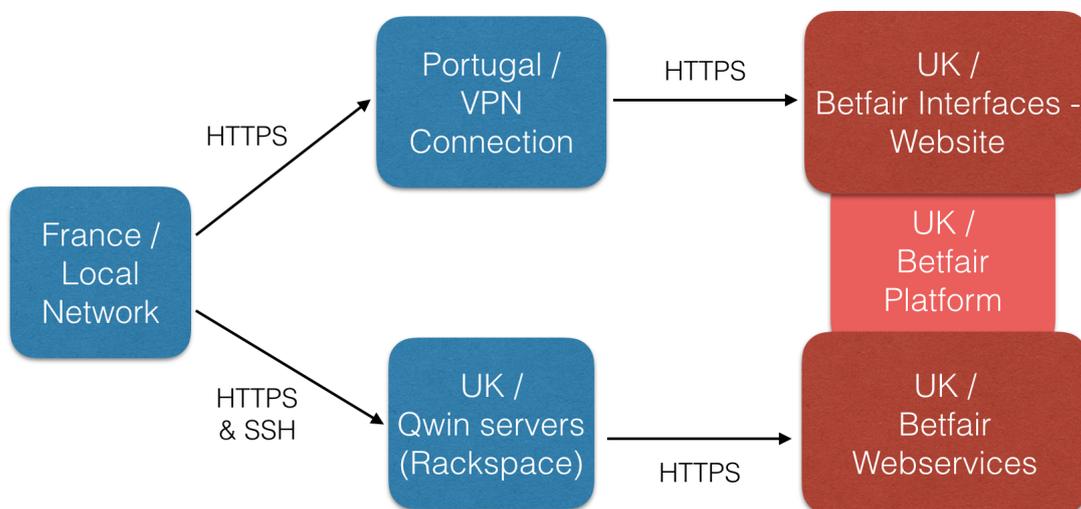


FIGURE 6.6 – High-Level-Design ; vue globale de l'architecture réseau

6.5 PHP et C++



FIGURE 6.7 – Logo PHP

PHP et C++ sont les langages que je maîtrise le mieux, c'est pour cela qu'ils ont été choisis.



FIGURE 6.8 – Logo C++

PHP a été privilégié pour sa rapidité de mise en oeuvre, ses performances, sa diversité et sa très bonne gestion des développements en mode POO (programmation orientée objet). Aussi, la librairie MySQL intégrée était un plus. Grâce à l'utilisation de MySQL Cluster, il n'a pas été utile de développer une classe de communication asynchrone avec le SGBD (comme cela aurait dû être le cas en cas d'utilisation de Postgres pour améliorer les performances). Concernant le C++, il a été utilisé pour des opérations de traitement de type Batch sur des ensembles de fichiers au niveau du système de fichier GlusterFS. Fortement intégré au système Linux et à l'utilisation des IO, et en utilisant la librairie Boost, des traitements rapides ont assuré les changements de formats des données avant injection dans le SGBD ou bien à des fins de calcul (création d'un nouveau score sur un ensemble de données statistiques collectées).

6.6 MySQL Cluster



FIGURE 6.9 – MySQL, le logo

L'expérience passée sur différents projets a été déterminante dans le choix du SGBD. Les technologies en lisse étaient Postgres, Cassandra, MongoDB, MySQL Cluster. Pour information, la gestion du système de fichiers distribué a été assurée par GlusterFS (mieux intégré au système que HDFS). Pour revenir au choix de la base de données, les fonctionnalités qu'apportent une base relationnelle tout en garantissant une scalabilité horizontale représentent les éléments principaux dans

la décision prise.

Aussi, j'avais une préférence pour les outils ne nécessitant pas de JVM, car le projet n'a pas été conçu en Java et je préférais éviter cette contrainte supplémentaire. Cela a donc été déterminant dans le choix de ne pas utiliser Cassandra.

De plus, nous disposons dans l'équipe d'une très bonne maîtrise du langage SQL, ce qui a permis de valider le choix de MySQL Cluster.

Dans ce genre de projet à budget réduit, il est indispensable de pouvoir prendre des décisions rapidement, de manière pertinente techniquement, sans disposer d'une période d'expérimentation trop importante pour envisager l'utilisation de produits inconnus. Cela veut aussi dire que dans un futur proche, en fonction des évolutions, un changement de technologie n'était pas exclu. C'est la formule de prise de décision : temps x compétence = productivité.

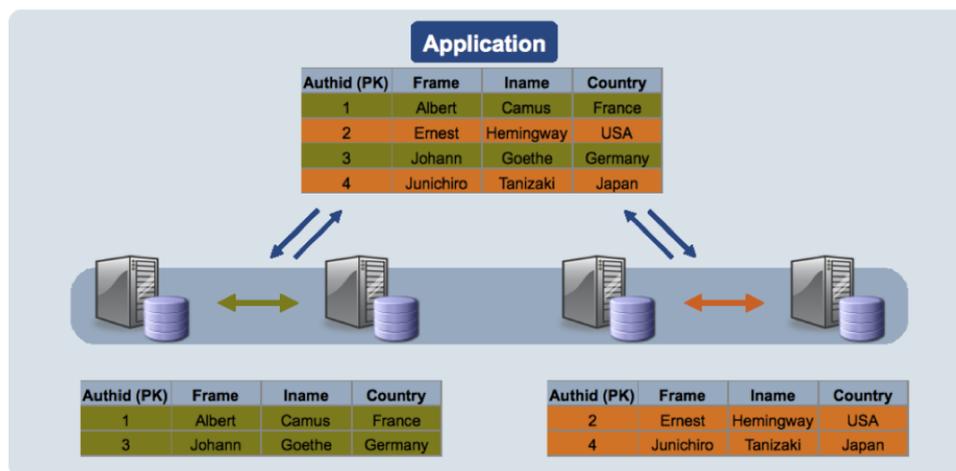


FIGURE 6.10 – Fonctionnement de l'auto-sharding MySQL

6.7 Apache et HHVM



FIGURE 6.11 – Apache, le logo

Incontestablement le serveur HTTP le plus répandu au monde.

Apache a directement été choisi et mis en oeuvre dans le cadre du projet pour sa souplesse de configuration, son support PHP, ses performances et son installation par défaut sur les environnements OSX.



FIGURE 6.12 – HHVM, le logo

HHVM (issu des travaux HipHop de Facebook) a été extrêmement utile lors des développements PHP multi-threadé.

Le simulateur a certainement été le programme le plus complexe à écrire et à tester.

Il a également été le programme le plus consommateur en ressource.

Afin de réduire les temps de calcul du simulateur, une conception orientée parallélisation des tâches a été entreprise.

Grâce à cela, il était possible de lancer le simulateur (Rewind) en mode multithread et en configurant le nombre d'exécution simultanée.

Ainsi, il devenait possible de lancer le simulateur en parallèle pour chacun des algorithmes et utiliser de manière optimisée l'ensemble des cœurs disponibles sur les processeurs de nos machines (lors du développement local ou sur la plate-forme de production).

A titre d'information, même si l'API HHVM autour des Posix Thread a donné des résultats très satisfaisants, la solution retenue fut de lancer en ligne de commande différents programmes / scripts PHP pour occuper l'ensemble des cœurs.

Cela a simplifié le développement (gestion des processus assurée par le système directement), tout en rendant le même service. Chaque exécution retournait ses résultats de calcul (voir § sur le simulateur) pour une présentation Web des tableaux et courbes de résultats par algorithme.

Chapitre 7

Spécifications

7.1 Stratégie de pari

Un grand nombre de paramètres de contrôle a été implémenté dans le système Qwin pour limiter les risques de pertes.

C'est ce que l'on peut appeler la stratégie, tout ce qui ne relève ni des automates de gestion des APIs ni des algorithmes de prédiction.

Cela peut être considéré comme une approche stratégique dans les méthodes de placement de paris, et l'on retrouve par exemple les conditions suivantes :

- La mise maximum par pari
- Les limites minimum et maximum des côtes
- Le montant maximum simultanément investi (30% de la trésorerie)

Beaucoup de ce qui aura été implémenté provient d'une étude personnelle sur les comportements à adopter dans le milieu du trading et également aux conseils issus du livre "Mastering Betfair"¹.

L'annexe "Trading Inplay, apprentissage et modèles prédictifs" reprend plus particulièrement l'ensemble des règles édictées tout au long du projet en fonction de découvertes successives, et le fonctionnement des marchés sportifs et les comportements observés.

1. Mastering Betfair de Pete Nordsted aux éditions Harriman House

7.2 Big Data

7.2.1 Collecte des données

Mapping sur les données

Comme expliqué au préalable dans la section "Etude" de ce mémoire, la base de tous les travaux de prédiction réside dans la capacité à disposer d'informations pertinentes, fiables et surtout en grande quantité.

Durant les 3 phases majeures du projet, différentes sources de données ont été utilisées (voir annexe "Grandes phases de l'approche fonctionnelle du projet").

- Prediday : Utilisation des données statiques issues des sites de statistiques et résultats sportifs (crawl par l'intermédiaire d'outils de collecte HTTP et conversion CSV avant injection en base)
- Paris avant match : Utilisation des données pré-citées, plus analyse des côtes Betfair récupérées au moment de l'ajout d'un nouvel évènement (nouveau match entre le PSG et l'OL le samedi suivant avec des côtes à un instant T en et en)
- Paris Inplay : Utilisation de toutes les informations en prédiction avant-match et utilisation des données dynamiques depuis Betfair concernant l'évolution des côtes durant la partie

Quelle que soit la provenance des données (sites de résultats et classement ou API Betfair), il a fallu procéder à un mapping des données pour injecter celles-ci à l'intérieur de notre base.

Type de données collectées

Voici l'ensemble des tables de la base de données MySQL, ces dernières comportent l'ensemble des champs utiles au mapping des données statistiques issues des sites de résultats et classement ainsi que de Betfair sur l'évolution des côtes et des volumes financiers négociés.

admin	ipref _r	pronodayusersmarketid
autoresults	log	proxy
bets	marketpoints	qwinrank
betsession	matches	qwinrewindgraph
bfbets	netstat	rangecontrol
bfdata	payment	rewindgraph
bfexchanges	placedbets	score
bfstats	popularity	secretbets
boardsession	pronoday	stats
illuminati	pronodaysession	users
ipexchanges	pronodaystats	
ipref	pronodayusers	

Détail ci-dessous sur la table des données de points de côte (collecte des données du marché en inplay chaque minute pour reproduction en simulateur) qui est devenue l'une des plus importantes du système pour les phases d'apprentissage :

```
--
-- Table structure for table 'marketpoints'
--

DROP TABLE IF EXISTS 'marketpoints';
/*!40101 SET @saved_cs_client = @@character_set_client */;
/*!40101 SET character_set_client = utf8 */;
CREATE TABLE 'marketpoints' (
  'marketid' varchar(50) NOT NULL,
  'date' int(11) NOT NULL,
  'sport' varchar(50) NOT NULL,
  'namea' varchar(50) NOT NULL,
  'nameb' varchar(50) NOT NULL,
  'namen' varchar(50) NOT NULL,
  'rateaback' float NOT NULL,
  'ratebback' float NOT NULL,
  'ratenback' float NOT NULL,
  'ratealay' float NOT NULL,
  'rateblay' float NOT NULL,
  'ratenlay' float NOT NULL,
  'pronostic' varchar(10) NOT NULL,
  'ip' tinyint(1) NOT NULL DEFAULT '0',
  KEY 'ratealay' ('ratealay'),
  KEY 'rateblay' ('rateblay'),
  KEY 'rateaback' ('rateaback'),
  KEY 'ratebback' ('ratebback'),
  KEY 'marketid' ('marketid'),
  KEY 'date' ('date')
) ENGINE=InnoDB DEFAULT CHARSET=latin1;
/*!40101 SET character_set_client = @saved_cs_client */;
```

Une configuration avancée des index en base de données a été mise en place afin d'optimiser les requêtes exécutées dans le simulateur comme dans les phases de trading en production.

Dans les premières versions de l'application, grâce au travail de collecte sur le site betexplorer.com par exemple, les prédictions se basaient sur un ensemble de données plus complet, comme le classement, le nombre de point, par type de compétition, et un grand nombre de score (score matchs extérieurs, score match à domicile, ...).

Ces informations étaient manipulées au-travers d'une classe de données statistiques par équipe (player).

La table bfstats regroupait ces informations statistiques.

```
class qwinStat
{
    private $_id;
    private $_date;
    private $_sport;
    private $_country;
    private $_competition;
    private $_surface;
    private $_player;
    private $_rank;
    private $_diff;
    private $_pts;
    private $_homerank;
    private $_homediff;
    private $_homepts;
    private $_awayrank;
    private $_awaydiff;
    private $_awaypts;
    private $_formrank;
    private $_formdiff;
    private $_formpts;
    private $_leagueDay;
    private $_totalDay;
    private $_pairId;
    private $_playerSide;
    private $_odd;
    private $_localization;
}
```

Le système dans sa dernière version ne se basait pratiquement plus que sur les données financières pour chaque match (montants disponibles sur le marché et côtes).

C'est pourquoi il s'agissait davantage d'un travail d'analyse statistique et probabiliste autour des règles financières que sur des analyses purement sportives.

Les membres de la classe ci-dessus constitués initialement la base des paramètres utilisés dans les algorithmes.

Compte-tenu du faible impact à grande échelle, des informations comme le type de championnat ou encore le nombre de journées de championnat déjà passées, ces données ont progressivement perdues du sens dans nos simulations.

En définitive, les côtes proposées par l'ensemble des traders sur l'Exchange Betfair donnaient des indications suffisantes, car les côtes elles-mêmes intégraient dans leur propre calcul les éléments statistiques que nous cherchions précisément à collecter au début du projet.

7.2.2 Traitement des données

Il y a eu plusieurs phases également dans le traitement des données.

La normalisation a été le premier besoin car entre des données d'un site de résultats à l'autre, la jonction parfois n'était possible que par le nom des équipes par exemple.

Outre l'encoding, certains fournisseurs modifiaient l'orthographe, ce qui nécessitait la création d'un module de mapping manuel entre les noms différents d'équipes identiques.

Cela s'est produit pour d'autres types de champs.

En terme de traitement, un programme C++ a permis la création d'une table de popularité.

C'est un des traitements qui a été mis en place durant la phase exploratoire sur les paramètres pouvant influés positivement sur les résultats de nos algorithmes.

Ce traitement de popularité calculait depuis le volume de données du moteur de recherche de Microsoft, Bing, apportait au système une indication sur le poids d'une équipe.

Même si l'on peut trouver une corrélation entre la popularité et le classement d'une équipe, ce paramètre n'a finalement pas été retenu pour nos prédictions.

Cela pour indiquer que divers traitements étaient appliqués sur les données afin d'en assurer la cohérence ou pour en extraire des informations nouvelles, après calcul (par exemple, calcul du nombre de victoire à domicile les 6 derniers mois).

Tous ces traitements étaient appliqués de 2 manières possibles : soit en batch avec écriture en base de données afin d'ajouter des nouveaux paramètres utilisables durant les phases temps réel de calcul de prédiction, soit directement durant le calcul des prédictions. L'impact sur les performances était l'élément déterminant pour la mise au point d'un nouveau traitement en mode batch dédié ou bien en mode ajout de méthode dans le processus même de calcul temps-réel.

7.3 Simulation

La capacité de reproduction des comportements de production sur la base stricte des données issues de la place de marchés sportifs Betfair a été la tâche la plus critique du projet.

Les détails d'implémentation seront précisés plus loin. L'accent dans cette partie du mémoire sera mis sur les objectifs à atteindre avec cet outil de simulation.

Le simulateur a pour objectif de se comporter comme le système de production, mais en apportant les nouveaux résultats issus des algorithmes en phase de test.

Cette recherche d'exactitude demande de conserver en base l'ensemble des événements sportifs, les évolutions dans le temps (évolution du classement par exemple) avec une date précise permettant une stricte reproduction.

Ce travail a permis de générer les résultats sur plusieurs mois voire plusieurs années, journée après journée, dans les différents sports.

En fonction des algorithmes sélectionnés pour la simulation, certains matchs étaient retenus avec des variations dans les prédictions et donc dans la balance financière finale suite aux résultats des paris.

C'est uniquement dans le but de disposer d'un simulateur performant qu'il aura fallu relever le challenge du stockage de données, car le volume que représente un stockage de l'ensemble des événements pour l'ensemble des sports, avec l'évolution des côtes et montants minute après minute depuis la création de l'évènement sur l'Exchange jusqu'au coup de sifflet finale, représente toute la valeur du projet.

Sans donnée, ou avec des données incomplètes, aucune simulation n'aurait été possible.

7.3.1 Prédiction avant-match

Les premières méthodes de prédiction étaient uniquement basées sur l'analyse des données à froid (classement, côte, nombre de buts marqués dans la saison, ...). Cette analyse pour chaque équipe permettait de déterminer l'écart entre celles-ci.

La première méthode de sélection d'un pari potentiel était donc basée sur la réduction du risque en fonction des paramètres ajoutés durant les expérimentations (en même temps ou de manière unitaire), dont voici un extrait :

- Le match se jouait-il à domicile ?
- Le nombre de journée dans le championnat
- La journée en cours
- Les informations disponibles annexes : match à l'extérieur, match à domicile
- Les côtes de chaque bookmaker pour 1N2 (en back et en lay, explication des termes plus loin dans le mémoire)

Les classes de traitements (statistiques par équipe et par évènement) et les de calculs (préparation des prédiction, placement de pari, algorithmes, ...) ont été conçus dès la phase de spécifications de manière à partager le même code entre la production et le simulateur.

Le simulateur a été imaginé et étudié de façon à ce qu'il reproduise de manière précise mais accélérée le fonctionnement de la production.

En rebouclant jour après jour sur les matchs à venir, le simulateur devait avoir pour seule objectif

de soumettre à un ou plusieurs algorithmes chacun des événements et, en fonction des résultats, assurer la consolidation pour chacune des journées.

Ces prédictions avant-match ont ensuite uniquement servi de filtre pour ne conserver que les événements sur lesquels lancer un trading Inplay.

Dans la section Conception et développement, le détail du fonctionnement du simulateur sera présenté pour comprendre sa mise en oeuvre à partir de cette spécification.

7.3.2 Prédiction inplay

Comme pour les prédictions avant match, les prédictions Inplay utilisaient les mêmes parties de code (classes, méthodes) que le programme de production, cela afin de ne pas risquer d'avoir à maintenir deux applications différentes et ainsi risquer d'avoir des écarts de comportement entre les paris en production et le simulateur. Cela aurait pu être fatal au projet très rapidement.

Comme en production, mais sans attendre le délai d'une minute entre chaque itération, le simulateur en mode inplay a en charge la récupération pour chaque prédiction des données du marché tout au long de la partie.

Le simulateur permet de modifier les algorithmes et de ce fait de chercher les optimisations possibles pour obtenir de meilleurs résultats.

7.3.3 Simulateur de résultats

Le simulateur devait alors, comme défini dans nos spécifications, et à partir d'un certain nombre d'options comme les algorithmes à tester, la période sur laquelle exécuter la simulation ou encore le montant à parier sur chacune des prédictions, générer le détail des opérations pour chaque journée ainsi que la synthèse des calculs : pourcentage de réussite par algorithme testé, nombre de paris placés, nombre de paris gagnés et perdus (et pour quels événements) et bien entendu balance financière suite aux calculs des gains et pertes sur la base des côtes historisées. Toute cela en n'omettant pas de produire un ensemble de logs facilitant le contrôle des traitements.

Ces logs ainsi que le rapport du simulateur (Rewind) sont des éléments indispensables à la validation des traitements pour assurer leur reproduction en production et l'absence d'erreur de calcul. Les annexes présentent un exemple de résultat Rewind.

7.4 Prédiction

7.4.1 Prédiction avant-match

Les prédictions avant-match ont constitué la première étape du projet. Après avoir collecté toutes les données statistiques (classement, nombre de points, ...) et conçu les premiers algorithmes basiques de type : "Si (Ecart de points entre Player A et Player B > 10) alors Parier Player B ne gagne pas " il est apparu simple de disposer d'un outil, sur la base des données disponibles dans le domaine publique, permettant d'obtenir 80% de réussite en terme de prédiction. La complexité avec ce genre de pourcentage est le suivant : Si l'on souhaite augmenter le pourcentage, cela réduit le nombre de pari possible. Si l'on veut augmenter le volume de pari et donc lisser le risque, cela réduit le pourcentage de réussite.

C'est donc le juste équilibre entre le type de pari, le volume de pari, le montant du pari et le pourcentage de réussite.

La performance d'un algorithme se mesurait donc sur sa capacité à délivrer un nombre important de paris avec un probabilité élevée de réussite.

Pour réussir un tel défi, il a fallu intégrer une nouvelle composante, le niveau des côtes. En effet, sur la base d'un pari avant-match fiable (mais avec des côtes faibles), il faut attendre les moments de doute durant la partie et le bon timing pour disposer de côtes intéressantes tout en conservant notre probabilité de réussite initiale. Les prédictions Inplay venaient de voir le jour dans les esprits. Il ne restait plus qu'à les intégrer au système de collecte et d'analyse.

7.4.2 Prédiction inplay

Toute la subtilité du projet a été de réaliser les meilleures prédictions, en ensuite de détecter le meilleur moment pour placer les paris, c'est à dire optimiser les côtes disponibles et cumuler prédictions et analyse haute-fréquence des marchés.

C'est uniquement durant le cours d'un match que les évolutions des côtes sont les plus prononcées.

Sur la base d'une prédiction à partir des éléments statistiques présentés dans la section précédente, et en fonction de l'évolution d'un match, les mécanismes d'analyse des côtes par observation des écarts permettaient de garantir les meilleurs côtes pour les meilleures prédictions.

Plus une prédiction avant match était bonne et laissait le système parier sur les côtes les plus rentables (ou les moins risquées), plus nous garantissions les gains. C'est là que se situait l'enjeu, tout en profitant de ces mécanismes d'analyse de côtes pour la collecte des données utiles à l'exécution des tests en simulateur. L'analyse Inplay a été probablement la phase du projet la plus complexe car le volume de donnée à traiter a été multiplié par 100 environ (analyse toutes les minutes).

Ce nouveau défi pour la collecte et le traitement a demandé une optimisation de notre infrastructure et de notre capacité de calcul. Mais de toute évidence a également permis d'intégrer une nouvelle finesse et optimisation financière dans les paris, garantissant une maîtrise des risques dans les possibles pertes pour chacun des paris.

7.5 Mode de pari

Le but dans ce mémoire n'est pas de lister exhaustivement les possibilités de paris sur des places de marchés sportifs.

Il existe une offre pléthorique de types de paris possibles et une complexité certaines dans le fonctionnement des distributions financières en fonction des côtes, en fonction des évènements et en fonction du nombre de pari en cours.

Je reprendrais ici les aspects génériques du fonctionnement de Betfair en particulier, et les grandes notions qui ont permis de mettre en place les automates.

Même si de nombreuses possibilités sont offertes par Betfair (Casino, Jeux, Turf, ...), nous nous sommes focalisés sur le football et le tennis, au niveau international et toutes compétitions confondues. Cela a fait partie des choix structurants pour ne pas se disperser et risquer de ne plus maîtriser notre environnement et nos algorithmes dédiés.

Le système était ainsi en mesure de parier jour et nuit sur des matchs de tennis ou de football à l'autre bout de la planète, avec des équipes et pour des compétitions qui m'étaient complètement inconnues.

7.5.1 Les différents types de marché

Dans le monde des paris, comme dans le monde de la finance, il y a une offre importante de produits sur lesquels il est possible de parier. Afin de maximiser les chances de réussite, la décision a été prise de travailler uniquement sur les paris de type 1N2. Ces marchés permettent de parier uniquement sur une victoire, une défaite ou un nul. Les autres marchés existants représentent une liste importante de possibilités, dont voici un extrait :

- Nombre de but sur la partie
- Nombre de but en première mi-temps
- Score à la mi-temps
- Différence de but
- ...

Pour éviter de devoir analyser et étudier un trop grand nombre de configurations et de statistiques, et surtout pour éviter de trouver des algorithmes sur des probabilités proches du hasard, nous avons donc restreint le type de marché au type 1N2 : victoire de A, nul, victoire de B (quel que soit le sport).

La seule subtilité qui subsiste ici vient du mode de pari.

En effet pour chaque cas d'un pari 1N2, il est possible de parier sur le fait qu'un évènement arrive, ou l'inverse (vrai ou faux). Cela nous donnera les possibilités suivantes :

- 1 gagne (pari de type 1N2 BACK)
- 1 ne gagne pas (pari de type 1N2 LAY)
- N est vrai (pari de type 1N2 BACK)
- N n'est pas vrai (pari de type 1N2 LAY)

- 2 gagne (pari de type 1N2 BACK)
- 2 ne gagne pas (pari de type 1N2 LAY)

7.5.2 Pari de type BACK

Comme expliqué succinctement ci-dessus, Betfair offre la possibilité de parier en BACK et en LAY.

La notion de BACK dont il est fait mention ici signifie qu'un évènement va se produire.

Il représente 1 cas sur 3 possibilités.

Il dispose également de gain plus faible car son fonctionnement est de type "Si VRAI, alors gain = mise x côte - mise".

En revanche, la perte est limitée à la mise si FAUX.

Il s'agit de la manière la plus courante de parier.

Sur les bookmakers traditionnels, en France également, il est courant de trouver ce type de pari.

Exemple : Match de football entre le Paris Saint Germain et Guingamp, au Parc des Princes. Le marché s'ouvre avec les évènements suivants, en BACK :

- Paris Saint Germain GAGNE : Côte 1.2 (forte probabilité, 80% environ)
- Match nul : Côte 2.1 (faible probabilité)
- Guingamp GAGNE : Côte 4 (très faible probabilité)

L'exemple présente bien la volonté de parier uniquement sur le fait qu'un évènement se produise ou non, avec une possibilité de faire un gain en fonction de la côte et du risque qui sera pris. C'est ce type de pari qui s'intitule un pari BACK.

Cela représente une chance sur trois au football et une chance sur deux au tennis (pas de match nul).

7.5.3 Pari de type LAY

La notion de LAY est plus subtile et vient directement compléter l'offre offerte par les paris de type BACK.

Il s'agit de la notion du "parier contre".

En utilisant le LAY, le calcul des gains est différent.

Le gain est systématiquement = mise x 2.

Ceci est une constante intéressante en terme de bénéfice.

Cependant, le calcul des pertes est différent et plus risqué.

Ici, c'est la côte qui déterminera une perte.

Un pari en LAY côte à 2.3 signifie donc : je parie que l'équipe X ne va pas gagner ou qu'il

y aura match nul. Nous avons ainsi 2 cas sur 3 pour que notre pari soit VRAI, avec la possibilité de doubler notre mise.

Par contre, si le pari est FAUX, donc que l'équipe X gagne, nous devons payer : $Perte = 2.3 \times \text{Mise}$, ce qui nous expose davantage que le pari BACK en terme de perte.

7.5.4 Notion de Bucket et fonctionnement de l'Exchange

Sur les places de marchés sportifs, il est possible d'analyser les buckets.

Les Buckets sont des plages d'investissement par côtes. Les Buckets sont donc les réservoirs financiers pour chaque marché.

Les équilibres se font entre les paris BACK, les paris LAY, les mises par côte. Chaque placement de pari venant augmenter un bucket en cours. En fonction du résultat d'un évènement, les Buckets sont ainsi redistribués aux gagnants, en rétribuant l'Exchange (Betfair dans notre cas) d'une commission en cas de victoire.

C'est donc en fonction de l'alimentation des Buckets et des écarts entre les évènements possibles que les côtes fluctuent.

Grâce à ce fonctionnement par bucket, il devient possible de parier sur différents étages.

Exemple de pari entre le Paris Saint Germain et Guingamp.

En LAY Guingamp, il y a par exemple :

- 2500 euros disponible sur le Bucket - Côte à 4
- 400 euros disponible sur le Bucket - Côte à 3.9
- 100 euros disponible sur le Bucket - Côte à 3.8

Imaginons que le système Qwin à cet instant décide de placer un pari en LAY sur Guingamp (Guingamp ne gagnera pas ou match nul), avec une mise d'un montant de 680 euros (calcul en fonction de la trésorerie disponible), il faudra (automatiquement) placer 3 paris :

- 1 pari de 180 - Côte à 4
- 1 pari de 400 - Côte à 3.9
- 1 pari de 100 euros - Côte à 3.8

Cela pour démontrer que les paris sont répartis par bucket.

Il faut alors chercher à miser sur les côtes les plus basses en LAY (et les plus hautes en BACK), le système étant capable de répartir une mise en plusieurs paris pour être certain d'utiliser toutes les sommes disponibles (mises BACK pour un pari LAY et inversement) sur chacun des buckets.

Ceci reste un exemple car en Inplay les paris sont plus fréquents (jusqu'à 1 par minute) et avec des montants moins élevés.

Voici une figure typique de représentation d'un marché sous Betfair, avec une vue simple des buckets par côte et des montants disponibles.

Irlande (ROI) - Montenegro		Égalé: EUR 56 292					Actualiser
<input type="checkbox"/> Paris live Scores en direct Statistiques sur les résultats							
<input checked="" type="checkbox"/> Back & Lay <input checked="" type="checkbox"/> Profondeur du marché							Plus d'options ▶
Sélections: (3)	100.4%	Back			Lay		99.4%
Irlande (ROI)	1.63 €1883	1.64 €906	1.65 €352	1.66 €451	1.67 €514	1.68 €1563	
Montenegro	7 €176	7.2 €400	7.4 €506	7.6 €220	7.8 €129	8 €163	
Match Nul	3.7 €523	3.75 €789	3.8 €16	3.85 €120	3.9 €632	3.95 €459	

FIGURE 7.1 – Betfair, Exemple de répartition par Bucket d'un événement

Des outils du commerce existent et permettent en temps réel de suivre l'évolution des buckets en fonction des paris réalisés.

Ces applications restent utiles uniquement pour des parieurs/traders manuels, qui positionnent des alertes aux moments qu'ils considèrent comme opportuns pour le placement d'un nouvel ordre.

Exemple d'applications d'aide aux paris : BetAngel, BetTrader, Geeks Toy, ...



FIGURE 7.2 – BetTrader, Exemple de suivi de l'évolution des côtes et des Buckets

La lecture du graphe permet de comprendre le fonctionnement de type "vase communicant" entre les paris de type BACK et de type LAY permettant à chaque partie de disposer d'un volume financier disponible en fonction des paris réalisés et par rapport à l'évolution des côtes.

Ainsi, en positionnant un pari à une certaine côte, cela rend disponible un pari opposé lui-même permettant l'évolution du marché.

C'est donc à celui qui proposera le meilleur pari, à la meilleure côte, et à celui qui donnera un ordre ensuite au meilleur moment, sur la meilleure côte.

Il est possible de proposer des paris à une côte donnée, avec un montant donné, sans que ces derniers ne soient encore disponibles sur le marché. Ils seront donc probablement retournés comme "Unmatched" ou "Partially Matched" car ils ne correspondront pas à la demande (côte trop éloignée des probabilités à un instant donné par exemple) ou que personne pour l'instant n'aura parié sur la totalité du montant du résultat opposé.

Ce comportement est expliqué sur de nombreux sites, et sur le site betting.betfair.com directement.

Dans le cadre de mes développements, je n'ai procédé qu'à des paris sur la base de buckets existants (montants disponibles à une côte donnée), afin de garantir leur réalisation (MATCHED ou PARTIALLY MATCHED) et une issue au pari.

En fonction de la rapidité des outils développés, il est arrivé que le pari ne passe pas ou plutôt reste en "UNMATCHED" car, le tant de la transaction, le bucket aura été pris par un autre trader (principalement durant les phases de paris "inplay").

Chapitre 8

Conception et développement

Le projet au total comprend plus de 95 000 lignes de code.

Il ne sera pas nécessaire de les reprendre ici dans leur ensemble, mais sera présenté les éléments les plus constitutifs du projet, le fonctionnement de modules principaux et les traitements les plus importants.

8.1 Présentation de l'API Betfair

Après les phases exploratoires autour des informations collectées sur les sites de statistiques (transformation CSV et stockage MySQL), il est apparu comme suffisant d'utiliser, dans leur intégralité, les informations fournies par l'API de Betfair pour l'ensemble des opérations des automates.

Betfair fournit des données en téléchargement libre sur les événements passés. Le problème rencontré avec ces jeux de données provient de l'absence des évolutions des côtes entre la création d'un marché et la fin de celui-ci (un marché est un type de pari pour un match, par exemple un pari 1N2 Back-Lay pour un match. Il existe des marchés sur le nombre de but marqué en première mi-temps également par exemple).

Les données statistiques de Betfair ne comprennent qu'une seule ligne par marché, et donc qu'une seule côte d'avant match.

Cela n'est pas suffisant pour permettre l'exécution des calculs du simulateur. En effet ce dernier, dans sa version finale, utilise les algorithmes d'avant-match pour la détection des événements à analyser, puis toutes les évolutions durant le match pour détecter l'ensemble des paris à placer durant celui-ci (inplay); et cela pour l'ensemble des algorithmes testés à chaque session du simulateur.

Fichiers CSV Betfair (historique des marchés et côtes) utilisés initialement pour les tests :

The Historical Data is provided "AS IS" without warranties of any kind (whether express or implied and including but not limited to warranties as to fitness for purpose, satisfactory quality and/or accuracy) and save as provided by statute and to the fullest extent permitted by law, Betfair excludes any and all liability whether in contract, tort, statute, equity or otherwise for any use of the Historical Data. The resale, distribution or the use of the service by another organization or individual is prohibited unless a specific agreement in writing with betfair exists. Additionally, we reserve the right to terminate or suspend this service at any time without prior notification.

Please note: If you have activated the country restriction functionality on your Betfair account, you may experience problems downloading data. We recommend you temporarily remove this restriction while downloading data. Please accept our apologies for any inconvenience. Due to the size of the data files, they are published in a compressed (.rar) format. Therefore, you may need to install an appropriate utility program (e.g. WinRAR) to open these files.

Horse racing data is separate from other data. The data is grouped by market/selection/odds/inplay. For example, we give a single record for the amounts traded on Ollie Magern in the Hennessy at odds of 8.0 prior to the race turning in-play.

The data given is as follows...

- EVENT_ID
- COUNTRY (horse racing only)
- FULL_DESCRIPTION
- COURSE (horse racing only)
- SCHEDULED_OFF
- EVENT
- ACTUAL_OFF
- SELECTION
- SETTLED_DATE
- ODDS
- LATEST_TAKEN (when these odds were last matched on the selection)
- FIRST_TAKEN (when these odds were first matched on the selection)
- IN_PLAY (IP - In-Play, PE - Pre-Event, NI - Event did not go in-play)
- NUMBER_BETS (number of individual bets placed)
- VOLUME_MATCHED (sums the stakes of both back and lay bets)
- SPORTS_ID
- SELECTION_ID
- WIN_FLAG (1 if the selection was paid out as a full or partial winner, 0 otherwise)

A list of Sports IDs is available [here](#).

We are aware of the following specific issues with the data

1. We occasionally have to re-settle a market. In these cases the volumes quoted may be inaccurate.
2. For Asian Handicaps and Markets such as Total Runs or Total Goals, there is no Settled Date on the market.
3. For Asian Handicaps, there is no indication of the handicap, although each line is a different event id.
4. With the introduction of Keep Bets at the end of 2008, some trades have subsequently been labelled as PE when they should in fact be IP. This is due the fact that the bets were matched in-play (IP) although they were placed pre-event (PE) using Keep Bets.
5. Italian Exchange bets are included in the 'Other' data, but this isn't specifically distinguishable between the existing UK/AUS Exchange data.
6. In the event that a market contains fewer runs than those which are being bet on to be placed, the original number of winners will be adjusted to be equal to the number of runners. Subsequently the market name will change automatically to reflect this. For example, a "4 TBP" would change to "3 TBP" in the event that only three runners took part in the race

Horse Racing Data	Other Data
2017 17 July 2017 - 23 July 2017 10 July 2017 - 16 July 2017 03 July 2017 - 09 July 2017 26 June 2017 - 02 July 2017 19 June 2017 - 25 June 2017 12 June 2017 - 18 June 2017 05 June 2017 - 11 June 2017	2017 17 July 2017 - 23 July 2017 10 July 2017 - 16 July 2017 03 July 2017 - 09 July 2017 26 June 2017 - 02 July 2017 19 June 2017 - 25 June 2017 12 June 2017 - 18 June 2017 05 June 2017 - 11 June 2017

Aujourd'hui Betfair a fait évoluer son site Historical Data afin de proposer l'ensemble des évolutions des côtes jusqu'à la fin d'un marché.

Le site <http://historicdata.betfair.com/#/home> présente cela.

Dans le cadre de notre projet, il a donc fallu constituer notre propre base de données d'historisation des évolutions des côtes par marché, car ces nouvelles archives plus complètes n'étaient pas encore proposées par Betfair.

Au démarrage seule l'API SOAP/XML était disponible puis j'ai migré vers la nouvelle API Betfair REST/JSON en créant un nouvel objet de traitement (bfexchange).

A partir de cette API, il était possible de réaliser l'ensemble des opérations de collecte d'évènements (match à venir), les côtes associées et leurs évolutions, ainsi que le placement des paris et le suivi du compte.

Voici une capture de la page d'accueil pour les développeurs utilisant l'API Betfair :

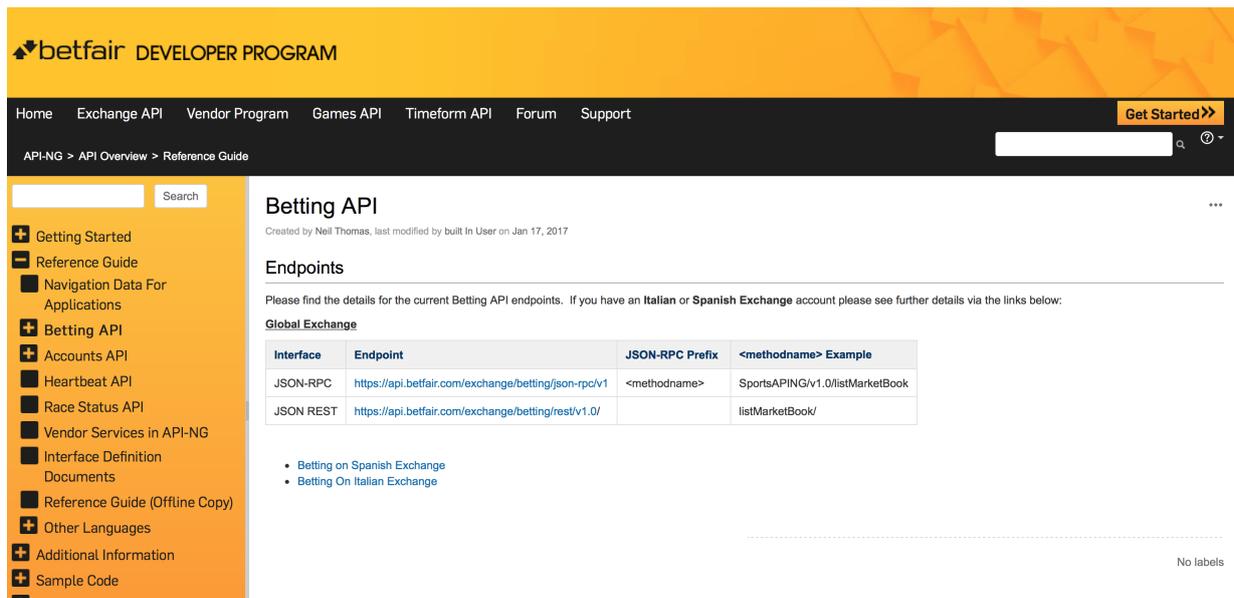


FIGURE 8.1 – Betfair, portail des développeurs

8.2 API Qwin - Webservice interne

Le module Webservice de Qwin permet l'accès aux données ainsi que la configuration du système depuis l'interface d'administration QwinBoard détaillée ci-après. Derrière ce Webservice, les bibliothèques partagées Qwin sont utilisées.

Liste des appels proposés par le Webservice :

```

case "refreshtomorrowstats" :
case "predict" :
case "version" :
case "sendsmsstatus" :
case "proxycheck" :
case "archivedb" :
case "refreshglobalstats" :
case "computebet" :
case "placebet" :
case "tradecontrol" :
case "marketpoints" :
case "getbotisalive" :
case "getnbgrabstats" :
case "getbetlist" :
case "getcustomerlist" :
case "getplacedbetlist" :
case "getstats" :
case "getaccountbalance" :
    
```

```
case "getsessionbetnumber" :  
case "getcurrentbetnumber" :  
case "getbetamount" :  
case "getsessionhardness" :  
case "getperiodvalue" :  
case "getrealbetvalue" :  
case "getqwinbotvalue" :  
case "gettradecontrolvalue" :  
case "getcontrolrateratio" :  
case "setcontrolrateratio" :  
case "setperiod" :  
case "setrealbet" :  
case "setqwinbot" :  
case "settradecontrol" :  
case "setsessionbetnumber" :  
case "setsessionhardness" :  
case "botalive" :  
case "getlog" :  
case "setlog" :
```

La simple lecture du nom des appels possibles de l'API Qwin permet d'appréhender l'utilisation qui a été faite du Webservice QwinAPI depuis les différents programmes clients (QwinBoard, robot de supervision et monitoring, robot de collecte et robot de placement de paris).

8.3 QwinBoard - Tableau de bord

L'interface d'administration permettait de disposer de la liste des évènements en cours, de la liste des paris placés, d'une vue globale de la plate-forme et des logs du système.

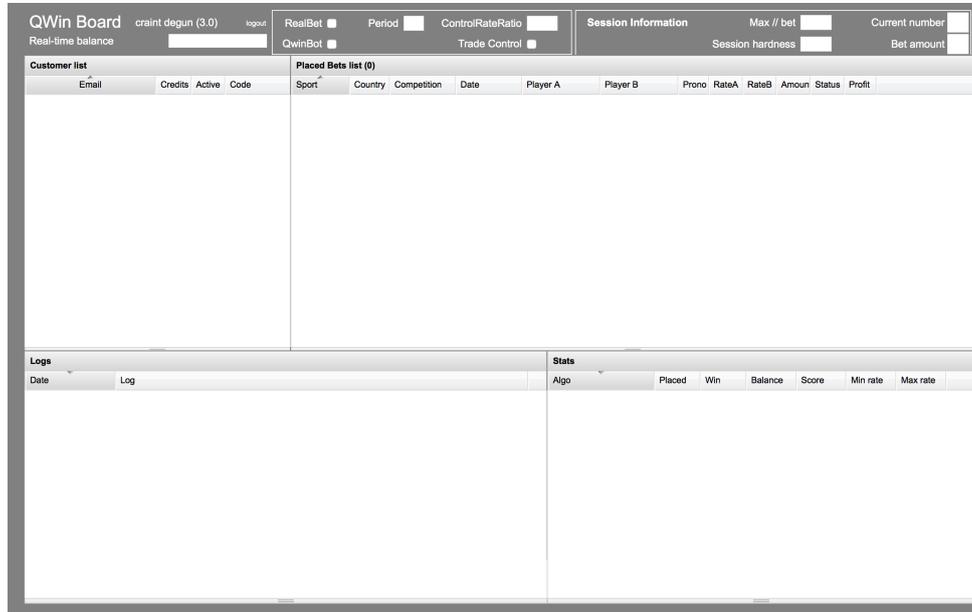


FIGURE 8.2 – Qwinboard, tableau de bord et administration

Le découpage en blocs (div) a permis tout d'abord de suivre les prédictions et la consommation des crédits par nos utilisateurs durant la toute première phase du projet (Prediday). Par la suite, QwinBoard a évolué, gardant son aspect principal, mais ne traitant plus que du système de trading Qwin et des évènements en cours et à venir.

Quelques captures de l'interface, en version mobile et en mode listing des paris en affichage détaillé, en page suivante.

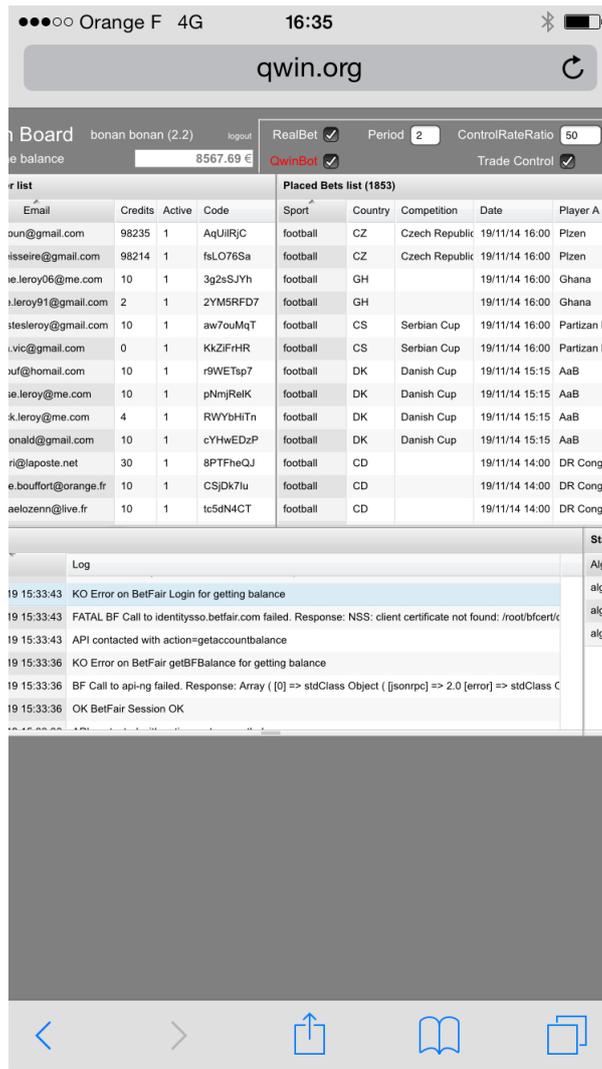


FIGURE 8.3 – Qwinboard, affichage depuis un téléphone mobile

info	ex	plot	1.116381609	GHG918E	23/11/14 18:00	football	SK	Slovakian Super Liga	Slovan Bratislava	Spartak Myjava	2L	algotPFFABS	0	S	17.15	30	30	0.
info	ex	plot	1.116381609	43773588024	23/11/14 18:00	football	SK	Slovakian Super Liga	Slovan Bratislava	Spartak Myjava	2L	algotPFFABS	0	M	23/11/14 18.42	11	11	0.
info	ex	plot	1.116381609	43775362726	23/11/14 18:00	football	SK	Slovakian Super Liga	Slovan Bratislava	Spartak Myjava	1L	algotPFFABS	0	M	23/11/14 19.25	21	21	0.
info	ex	plot	1.116381609	43775406216	23/11/14 18:00	football	SK	Slovakian Super Liga	Slovan Bratislava	Spartak Myjava	1L	algotPFFABS	0	M	23/11/14 19.26	22	22	0.
info	ex	plot	1.116381609	43775452696	23/11/14 18:00	football	SK	Slovakian Super Liga	Slovan Bratislava	Spartak Myjava	1L	algotPFFABS	0	M	23/11/14 19.27	22	22	0.
info	ex	plot	1.116381609	43775502887	23/11/14 18:00	football	SK	Slovakian Super Liga	Slovan Bratislava	Spartak Myjava	1L	algotPFFABS	0	M	23/11/14 19.28	8	8	0.
info	ex	plot	1.116381609	43775510039	23/11/14 18:00	football	SK	Slovakian Super Liga	Slovan Bratislava	Spartak Myjava	1L	algotPFFABS	0	M	23/11/14 19.28	2	2	0.
info	ex	plot	1.116381609	43775517356	23/11/14 18:00	football	SK	Slovakian Super Liga	Slovan Bratislava	Spartak Myjava	1L	algotPFFABS	0	M	23/11/14 19.28	5	5	0.
info	ex	plot	1.116381609	43775524780	23/11/14 18:00	football	SK	Slovakian Super Liga	Slovan Bratislava	Spartak Myjava	1L	algotPFFABS	0	M	23/11/14 19.28	7	7	0.
info	ex	plot	1.116381609	43775556436	23/11/14 18:00	football	SK	Slovakian Super Liga	Slovan Bratislava	Spartak Myjava	1L	algotPFFABS	0	M	23/11/14 19.29	22	22	0.
info	ex	plot	1.116381609	43775605167	23/11/14 18:00	football	SK	Slovakian Super Liga	Slovan Bratislava	Spartak Myjava	1L	algotPFFABS	0	M	23/11/14 19.30	23	23	0.

FIGURE 8.4 – Qwinboard, tableau détaillé des paris

8.4 BFStats - Module de collecte des évènements

Toutes les 10 minutes, une récupération globale de tous les évènements sportifs est effectuée afin de permettre au programme de calcul des prédictions de se baser sur les marchés à venir. De nombreux paramètres sont disponibles dans ce module, comme la durée entre le moment de collecte et le début de l'évènement (par exemple tous les évènements entre maintenant et une semaine).

8.5 QwinBet - Module de prédiction et placement des paris avant match

Grâce au module QwinBet, exécuté également à l'aide d'une entrée dans la Crontab, l'ensemble des évènements à venir était analysé puis filtré grâce aux algorithmes de détection de matchs à fort potentiel de paris dans la phase d'inplay (du coup d'envoi à la fin du match).

QwinBet disposait d'un ensemble de paramètres (montant à parier, délai minimum entre l'analyse et le coup d'envoi, ...).

QwinBet, pour son exécution, utilisait donc en entrée :

- %lPronoParams->myBetAmount : le montant à parier
- %lPronoParams->myMinBetRate : la côte minimum pour placer un pari
- %lPronoParams->myMaxBetRate : la côte maximum pour placer un pari
- %lPronoParams->myMinRankCompEntries : le nombre minimum d'entrée en base par compétition pour réaliser une prédiction convenable
- %lPronoParams->myGlissingWindow : la période de statistiques sur laquelle réaliser les calculs servant à la prédiction
- ...

8.6 IPControl - Module de collecte et de placement des paris "inplay"

La méthode de pari en inplay est devenue dans le système le module principal de trading. En effet, après le filtrage des matchs en avance de phase, le module IPControl, en plus de collecter la totalité des états des buckets de chacun des évènements (dans le but de tester de nouveaux algorithmes grâce au simulateur par la suite), avait la responsabilité d'exécuter les algorithmes de production tout au long de chacun des matchs retenus et de placer, au rythme maximum d'une fois par minute et avec des limites dans les montants à parier, les paris LAY (BACK possibles mais moins intéressants) issus des prédictions en temps-réel.

Le module était découpé en 2 scripts, l'un pour la collecte des points de marché et l'autre pour le trading à appliquer.

8.7 QwinRank - Module de ranking des players

Dans la phases de prédictions avant-match, beaucoup de paramètres ont été utilisés afin d'être en capacité de filtrer au mieux les évènements pour lesquels le trading inplay devait s'appliquer. QwinRank est le module en exécution de type batch, qui permet le rafraîchissement de la table SQL interne de gestion du classement de chacune des équipes ou joueurs (player dans le vocabulaire du betting).

8.8 Popularity - Module de calcul de popularité

De même nature que le batch QwinRank, ce module avait pour objectif la mise à disposition de statistiques concernant les players. La popularité a été utilisée dans les premières phases du projet, mais ne s'est pas avérée être en définitive un élément pertinent dans l'amélioration des calculs de probabilité.

La popularité était basée sur un grand nombre d'informations, comme le nombre de résultats issus d'une recherche sur une équipe sur le moteur de recherche Bing©de Microsoft™.

8.9 Rewind - Module de reporting et simulation

Ce module, celui sur lequel j'ai probablement passé le plus de temps à lancer des tests d'optimisation sur les algorithmes à l'essai, a été développé avec le plus grand soin en concevant un ensemble de classes de type machine à état, permettant de lister l'ensemble des évènements passés et de reproduire à l'identique les évolutions du marché et donc des potentiels paris que le système proposait.

Un rapport Rewind comprenait 5 sections :

- Le paramétrage du simulateur et les options disponibles
- Un récapitulatif des options en cours (2 algorithmes testés, montant 30 euros par pari)
- Des logs (mode debug) sur des points spécifiques en cours de contrôle
- Le déroulé jour après jour des prédictions et placements de paris avec l'évolution de la balance associée
- Le bilan pour chaque algorithme

Pour chaque journée de paris, le simulateur présentait pour chacun des algorithmes :

- La date
- Le nombre de paris placés
- L'état de la balance
- La couleur verte si la balance a augmenté, rouge dans le cas contraire
- L'exposition (montant total investi durant la journée, risque de perte maximum)

Rewind, le simulateur, a produit un nombre incalculable de tableau et rapport permettant de qualifier le potentiel de chacun des algorithmes.

Un exemple de sortie Rewind est intégré aux annexes (Annexes "Rewind, exemple de résultats en sortie du simulateur").

Chaque journée disposait également du détail des paris placés et des montants associés (sur passage de la souris dans les cellules TD du tableau), également dans les annexes.

Les graphiques générés en plus des tableaux donnés une rapide lecture de la tendance de l'algorithme concerné.

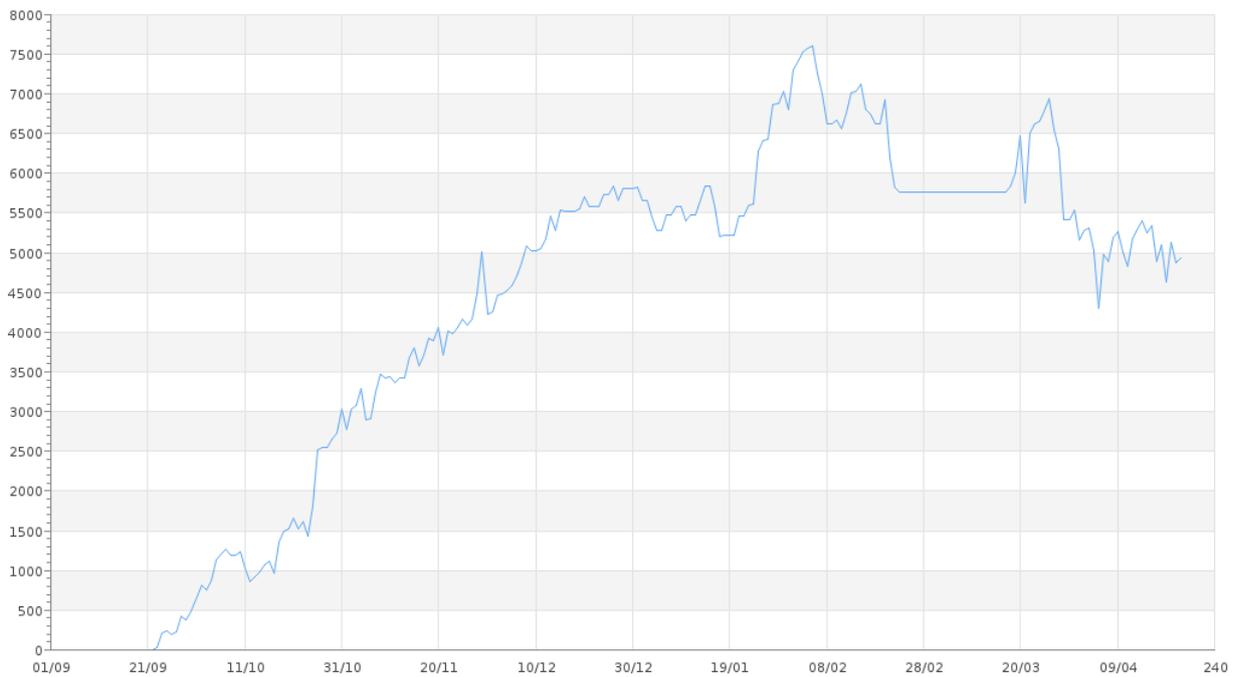


FIGURE 8.5 – Rewind, génération de graphique sur l'évolution de la balance financière

8.10 Module de gestion de bibliothèques internes et de la base de données

Les bibliothèques internes et partagées à l'ensemble des modules sont principalement :

- Bibliothèque de communication avec l'API Betfair
- Bibliothèque de conception des algorithmes
- Bibliothèque de communication avec le cluster MySQL
- Bibliothèque de gestion des alertes
- Bibliothèque de d'envoi d'email et de sms
- Bibliothèque de récupération des statistiques
- Bibliothèque de gestion de l'identification et des sessions
- Bibliothèque de gestion de la trésorerie

Ci-dessous la liste des bibliothèques développées dans le cadre de la mutualisation des classes transverses du système Qwin. Y figurent les classes pour l'ensemble des fonctionnalités comme l'envoi de SMS dans les modules de supervision, la gestion de la trésorerie, les algorithmes de production, la classe en charge du parsing HTML des sites de statistiques ou encore des outils comme la classe de gestion des user-agents pour ne pas être *blacklisté* par les fournisseurs et assurer un changement d'identité (user agent, proxy).

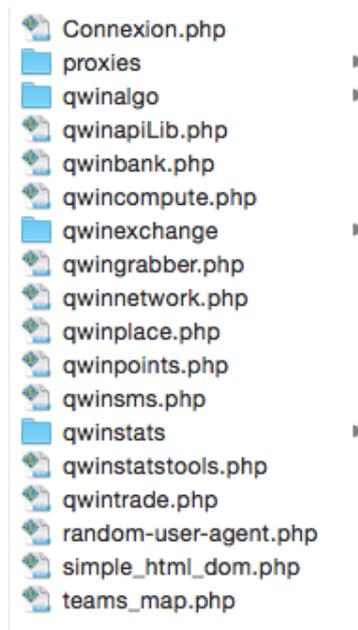


FIGURE 8.6 – QwinLib, bibliothèques transverses du système Qwin

La classe de gestion de l'API Betfair a probablement été la plus complexe à développer. Cela certainement car elle a nécessité la gestion de certificat, d'une authentification forte à base de token délivré par Betfair et une API JSON/RPC très bien documentée mais nécessitant de notre part une couverture à 100% des cas d'erreur afin de maîtriser totalement les opérations financières exécutées sur la place de marché durant les phases de trading.

Voici un exemple d'appel à l'API Betfair. La fonction *sportApingRequest* est une méthode privée de la classe de traitement Qwin, garantissant un appel cURL sécurisé et contrôlé.

```

function getMatches(&$pMatchArray)
{
    // "2013-07-20T00:00:00.00Z"
    $lFrom = "";
    $lTo = "";
    // Get period from DB
    $lResult = mysql_query("select period from admin");
    $lRow = mysql_fetch_row($lResult);
    $lPeriod = 1;
    if($lRow[0] > 0)
    {
        $lPeriod = $lRow[0];
    }

    // If period=2, means today and tomorrow (in fact time() to 48h from now)
    // If period=3, means from now to now+62h
    for($i=0; $i < $lPeriod; $i++)
    {
        // $lFrom = "2013-07-21T00:01:00.00Z";
        // $lTo = "2013-07-22T00:01:00.00Z";

        $lNextTime = mktime(date("G"), date("i"), date("s"), date("n"),
            date("j")+ $i, date("Y"));
        $lFrom = date("Y-m-d", $lNextTime)."T".date("H:i:s.00", $lNextTime)."Z";

        $lNextTime = mktime(date("G"), date("i"), date("s"), date("n"),
            date("j")+ $i+1, date("Y"));
        $lTo = date("Y-m-d", $lNextTime)."T".date("H:i:s.00", $lNextTime)."Z";

        doLog("BF getMatches from ".$lFrom." to ".$lTo);

        // Remove "turnInPlayEnabled":"true", from filter .... 25/08/2013
        // '{"filter":{"eventTypeIds":["1","7522","998917","7511","7524"]},
        // 05/11/2013 : Foot only, have to bench others sports ...
        $params = '{"filter":{"eventTypeIds":["1"],
            "MarketBettingType" : "ODDS",
            "marketTypeCodes" : ["MATCH_ODDS"],
            "turnInPlayEnabled":"true","inPlayOnly":"false",
            "marketStartTime":{"from":"' . $lFrom . '", "to":"' . $lTo . '"}},
            "marketProjection":
            ["COMPETITION", "EVENT", "EVENT_TYPE", "MARKET_START_TIME", "RUNNER_DESCRIPTION"],
            "sort":"FIRST_TO_START",
            "maxResults":"200"}';

        $jsonResponse = $this->sportsApingRequest($this->myAppKey,

```

```
        $this->mySessionToken, 'listMarketCatalogue', $params);

    if($jsonResponse != false)
    {
        foreach($jsonResponse[0]->result as $lMatch)
        {
            array_push($pMatchArray, $lMatch);
        }
    }
}

return;
}
```

Dans cette méthode, il est possible de comprendre le fonctionnement de l'API Betfair et la manière de configurer chaque requête.

Il s'agissait dans cet extrait du code Qwin de récupérer l'ensemble des matchs à venir (coup d'envoi pas encore donné) sur une période donnée (de maintenant à dans une semaine généralement), et pour lesquels le trading inplay sera activé (essentiel dans notre stratégie de placement des paris).

Chapitre 9

Tests et Optimisations

9.1 Tests unitaires

Avant chaque livraison d'une nouvelle version et son packaging, un ensemble de tests unitaires étaient exécutés dans le but de garantir le niveau fonctionnel du système Qwin.

Ce niveau fonctionnel représente une partie de la non-régression, mais également la cohérence de l'application.

Si le type de retour d'une fonction diffère d'une version à l'autre ou bien un paramètre d'appel à fonction, c'est le système dans son ensemble qui est affaibli et risque de se comporter de manière non-maîtrisée, ce qui provoquerait potentiellement un risque de perte financière directe.

Voici la liste des fichiers de tests unitaires appelés durant la phase de build et packaging d'une nouvelle version.

Comme dans tout système de tests unitaires, des mocks étaient développés afin de permettre à chaque classe et chaque fonctionnalité de disposer des entrées attendues dans le cadre de l'analyse de leur sortie.

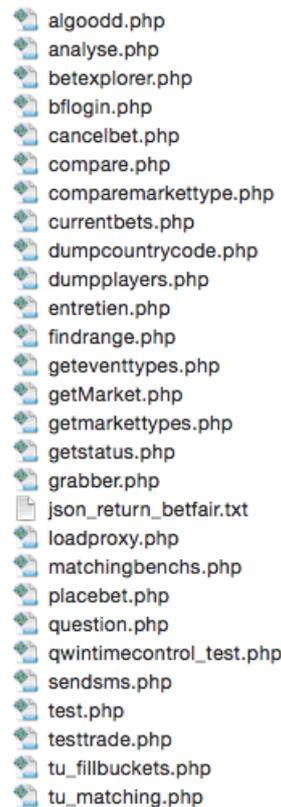


FIGURE 9.1 – QwinTU, liste des scripts de test

9.2 Tests de non-régression et logs

La non-régression est une phase indispensable pour tout projet informatique. C'est la garantie d'un passage en production sans encombre. Cette phase de non-régression dans le cadre du projet Qwin était assurée par la configuration d'un mode "Virtual Bet" sur l'ensemble des modules. Cela me permettait d'observer directement sur des données et des évènements de production comment le système se comportait et aurait placé ses paris. Toute écriture et ordre Betfair étaient donc contournés, les logs du système largement analysées pour valider le fonctionnement attendu des modules et des algorithmes.

Les modules du système Qwin généraient leur propre fichier de logs, tout en respectant un format unique, produit par la librairie dédiée de gestion des logs. Cette méthode permettait, en suivant les identifiants uniques des évènements, de croiser toutes les informations des fichiers de logs et présenter les séquences de traitements, facilitant ainsi toutes les phases d'investigation.

Chapitre 10

Mise en production

10.1 Pré-production

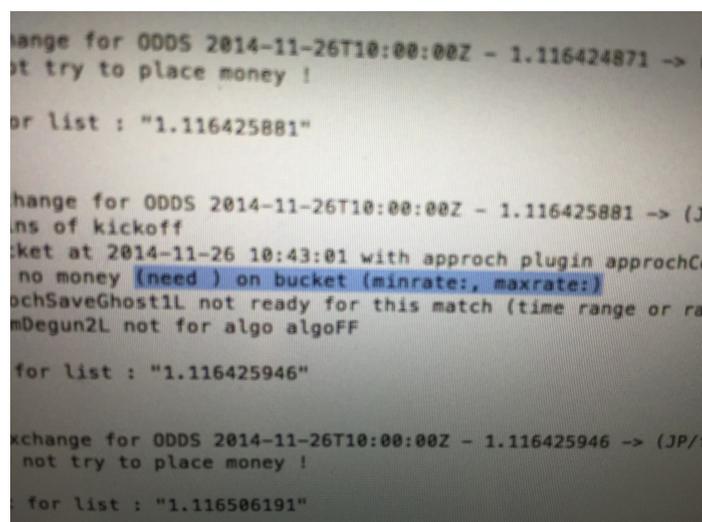
La pré-production se réalisait directement sur les serveurs de production. En effet, grâce à des mécanismes de tirs à blanc évitant tout risque de placement de pari en erreur, il était possible de déployer le nouveau code validé sur l'environnement cible.

Après une phase de vérification du bon comportement, le mode nominal de fonctionnement était réactivé.

Cela afin de garantir un circuit court, de même type que ce que nous trouvons aujourd'hui en entreprise sous la forme du DevOps. Depuis le commit jusqu'à la livraison, un ensemble de mécanismes assure la validation de la nouvelle version.

Le DevOps était appliqué pleinement d'un point de vue automatisation car j'assurais l'ensemble des phases de *release management*, de suivi du versionning, de packaging, d'intégration et de déploiement.

L'analyse des fichiers de logs et par conséquent du comportement du système permettait de valider les nouvelles versions (Voir extrait des logs en annexe "Extrait des logs du système Qwin").



```
change for ODDS 2014-11-26T10:00:00Z - 1.116424871 -> (
at try to place money !

for list : "1.116425881"

change for ODDS 2014-11-26T10:00:00Z - 1.116425881 -> (JP
ns of kickoff
cket at 2014-11-26 10:43:01 with approach plugin approachCo
no money (need ) on bucket {minrate:, maxrate:}
chSaveGhost1L not ready for this match (time range or rat
eDegun2L not for algo algoFF

for list : "1.116425946"

change for ODDS 2014-11-26T10:00:00Z - 1.116425946 -> (JP/f
not try to place money !

for list : "1.116506191"
```

FIGURE 10.1 – Qwin log, outil d'investigation

10.2 Real betting

A chaque nouvelle version, contenant de nombreuses évolutions ou simplement un bugfix, et après avoir réalisé les tests au simulateur, les tests unitaires et de non-régression, une période d'observation en production était effectuée.

Cette phase d'observation permettait d'exécuter l'ensemble des programmes, et de vérifier en base de données et dans les logs applicatifs que les traitements correspondaient bien aux comportements attendus.

A partir de ce moment, une fois la validation complète réalisée, l'activation du Real Betting se faisait par un paramétrage en configuration, permettant de lancer les ordres de placement de paris auprès de l'opérateur Betfair.

Chapitre 11

Livraison et sauvegarde

11.1 Livraison

La livraison s'effectuait par l'intermédiaire de scripts shell et de synchronisation SSH via l'utilitaire SCP.

Cela permet un déploiement simple vers l'ensemble des serveurs et vers l'environnement choisi.

```
#!/bin/sh
# $1 can be dev preprod or prod

source ./cleanenv.sh # remove cache files, pictures, local logs before create package

tar -czvf $1.tar.gz $1/

scp -P 7070 $1.tar.gz root@qwin.org:/home/qwin/
ssh -p 7070 root@qwin.org "cd /home/qwin/; ./deploy."$1".sh"
```

L'utilisation du port 7070 émane d'une volonté de ne pas utiliser de ports standards pour l'exploitation de nos solutions, afin de diminuer le risque d'intrusion par scan des ports TCP et UDP des serveurs (en plus des règles Firewall).

Une fois le package déposé sur le serveur, la commande `./deploy.[preprod|prod].sh` était exécutée pour installer la nouvelle version.

11.2 Sauvegarde

Sur le même principe de gestion via scripts shell, il a fallu concevoir des mécanismes de transfert des applications, des logs ainsi que de la base de données.

Grâce à notre VPN et un serveur local dédié, les sauvegardes étaient assurées chaque nuit avec un système de rotation type logrotate sur une semaine pour ne pas saturer la machine de stockage. Les données ainsi que la base de données étaient archivées à l'aide de la commande tar et compressées avec l'option gzip (paramètres utilisés : `czvf`).

Chapitre 12

Recherche d'algorithmes

12.1 Phase exploratoire

La phase exploratoire consistait en l'étude sur tableau blanc et sur papier des éléments constituant les phases de prédiction (élaboration des modèles prédictifs).

En s'intéressant tout d'abord de très près à chacun des sports pour déduire les informations pertinentes, et ensuite par l'apprentissage des comportements des marchés financiers, les algorithmes ont successivement été améliorés, modifiés, réinventés et bien entendu qualifiés.

La qualification des nouveaux algorithmes se faisait au-travers du simulateur Rewind sur la période la plus large possible.

Les courbes et tableaux de résultats obtenus nous permettaient de travailler sur les nouvelles pistes et axes de calculs.

Des données manquantes étaient donc collectées ou bien générées à partir de données existantes, des essais étaient réalisés, et toute la complexité résidait dans la capacité à produire un algorithme qui soit à la fois joueur (c'est à dire qui propose un nombre suffisant de paris), et performant (qui dispose d'un niveau satisfaisant de réussite dans ses prédictions $> 80\%$).

12.2 Essais en simulateur

Chaque nouvel algorithme devait faire l'objet d'une validation, sur une période de temps la plus longue possible (jusqu'à 5 ans).

Compte-tenu du temps d'exécution, il a été utile de concevoir un simulateur en mode multithread et ainsi utiliser pleinement la puissance machine comme on le voit sur la capture ci-dessous où le simulateur exploite chaque coeur de la machine sur ses traitements.

Comme le montre le graphique Rewind en annexe, représentant l'évolution financière d'un des algorithmes en phase de test, la mise au point d'une approche à forte croissance représentait un succès. Cependant, et ce fût le principal problème auquel il a fallu faire face, malgré une base de données de plusieurs milliers de résultats sur plusieurs années, le travail d'optimisation des algorithmes amenait souvent à la réalisation d'un résultat sur mesure par rapport à l'exploitation des données existantes.

Le résultat en production ne reproduisant que rarement la courbe du simulateur, il convenait

de conclure qu'il résultait du travail en simulateur uniquement une capacité à trouver l'équation optimale pour en extraire les meilleurs performances a posteriori.

L'enjeu se situait donc dans la capacité à concevoir des algorithmes performants une fois passés en production.

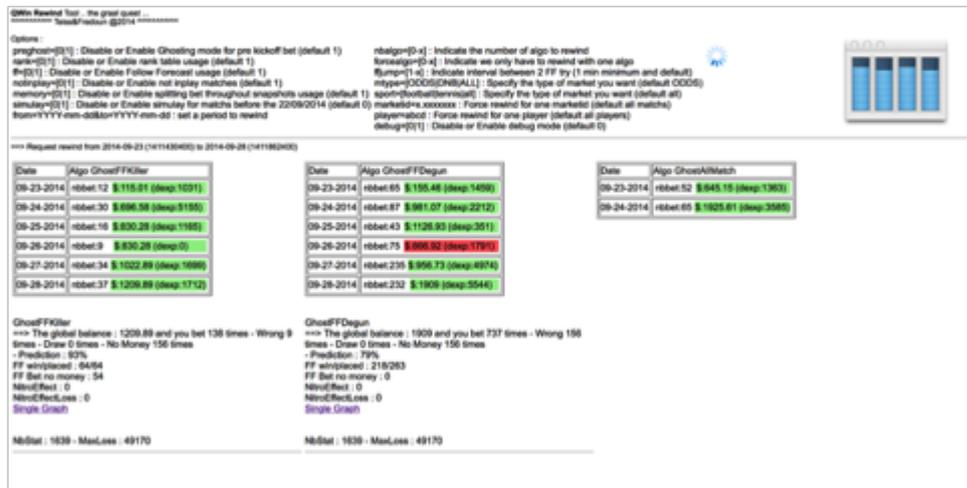


FIGURE 12.1 – Rewind Multithread, capture de l'utilisation de l'ensemble des coeurs du CPU

12.3 Passage en production

Comme détaillé plus haut, les mises en production permettaient un mode "Virtual Bet" qui assurait la mise en service de nouveaux algorithmes pour une analyse approfondie du comportement avant l'activation réelle des placements de paris une fois les nouvelles prédictions validées.

Chapitre 13

Bilan financier

Il est important dans tout projet de définir le cadre d'investissement (le besoin initial) ainsi que les perspectives de croissance.

Le lancement d'un projet est ainsi le moment où les protagonistes sont en mesure de présenter :

- Une capacité à investir du temps dans le projet
- Des compétences pour l'accomplissement du projet
- Une capacité à investir financièrement dans le projet
- Les prévisions de croissance au travers d'un business plan
- Un prototype opérationnel démontrant la viabilité du projet

Il est différent de parler de projet et de parler d'invention. Dans le cas présent, il s'agit foncièrement d'un projet. Je ne vais pas dans ce chapitre aborder l'ensemble des points listés ci-dessus mais uniquement me concentrer sur les aspects financiers.

C'est en cela que je parle de projet, car il y a bien une dimension financière, un objectif de rentabilité et de *profitabilité* annoncé dès le démarrage.

13.1 Budget initial

Comme indiqué dans la section 'Exigences', le budget initial du projet était de 100 000 euros. Cette somme a été établie sur la base des 3 besoins du projet.

Le besoin premier était la rémunération. Compte-tenu du fait que le projet a été réalisé par 2 personnes, à temps plein, une capacité de financement équivalent à un an de salaire pour 2 personnes représentait ce besoin, soit la somme de 72 000 euros (3000 euros par mois par personne sur 12 mois).

Le projet a duré 2 ans, mais la première année, la rémunération personnelle était assurée par d'autres sources de revenus.

Le deuxième besoin est celui de l'investissement logistique : bureau, ordinateurs, serveurs, hébergement. Cette somme se limite à 8 000 euros car nous disposions déjà de nombreux équipements et le loyer est resté très accessible. Enfin, le troisième besoin en financement était la trésorerie pour l'exécution des paris sportifs. Cette somme était portée à 20 000 euros.

C'est sur cette base de 100 000 euros que le projet a pu être réalisé dans son ensemble.

13.2 Gestion de la trésorerie

J'évoque dans cette section la trésorerie au sens du montant disponible sur la place de marchés (cash-flow BetFair).

Après avoir défini les règles, celles-ci ont été programmées pour qu'elles s'appliquent automatiquement dans toutes les phases de gestion des fonds des automates.

Liste de certaines des règles édictées :

- Ne jamais jouer sur la base d'une intuition, d'une émotion : assurer par la mise en place de 'bot'
- Ne jamais investir en même temps plus que 30% de sa trésorerie totale
- Ne jamais faire un pari d'un montant supérieur à 10% de la trésorerie
- Ne jamais faire de pause dans le placement de paris, toujours maintenir un rythme pour lisser dans le temps les courbes

Pour chaque nouveau pari possible, nous déterminions *programmatically* le montant du pari en fonction du nombre de pari en cours à un instant T ainsi que d'autres paramètres utiles.

Extrait de code sur l'appel à la fonction de calcul du montant à parier :

```
// Compute bank in order to know the actual betamount
$1BetAmount      = computeBank($1Balance, $1SessionBetNumber, $1CurrentNbet,
                               $1PreviousBetAmount, $1Hardness, $1TradeControlStatus);
```

Les détails de la stratégie de placement de paris sont détaillés en annexe "Trading Inplay, apprentissage et modèles prédictifs".

Chapitre 14

Conclusion

14.1 Synthèse financière

D'un point de vue financier il y a eu une gestion intelligente et maîtrisée du cash-flow. Les règles sont simples mais indispensables : Ne jamais parier plus que 30% de sa trésorerie en même temps et ne jamais arrêter la machine en cours de cycle.

A partir de cela, j'ai pu gérer et limiter la prise de risques.

Avec une trésorerie Betfair au meilleur de l'activité de 80 000 euros (sur un investissement initial de 20 000 euros), j'ai pu observer des périodes de forts gains, des périodes de gains relatifs et des périodes de fortes pertes.

Le manque de profondeur, uniquement 6 mois, de la base de données finale (celle qui permettait de disposer de l'ensemble des informations utiles aux algorithmes les plus pertinents) n'aura pas permis de franchir la frontière qui mène du projet à l'entreprise.

J'ai alors préféré arrêter, de concert avec mon partenaire, en me laissant la possibilité de collecter des données pour un traitement sur un ensemble plus grand de données a posteriori.

Les jeux de données statiques de plusieurs années proposés gratuitement par Betfair, malheureusement, ne me procuraient pas l'intégralité des informations utiles.

La limitation a donc été financière, non pas par perte importante, mais par l'effet induit du manque de temps.

Il n'y aura eu de perte que la rémunération durant les mois de projet et une réalisation des travaux à temps plein.

Avec un investissement initial de 100 000 euros, j'ai donc été en mesure de récupérer les sommes investis dans la trésorerie Betfair (20 000 euros) ainsi que les investissements matériels (8 000 euros).

Seule la partie rémunération (72 000 euros) restera une perte sèche.

Comme souvent dans les projets, il est possible de récupérer la mise sur les investissements, mais rarement sur la rémunération, le centre de coût le plus critique.

N'ayant pas réussi à générer des bénéfices, mais uniquement à rester à l'équilibre dans la balance des paris, la rémunération reste le poste de dépenses qu'il n'a pas été possible de récupérer ni de rendre profitable avec des dégagements de bénéfices.

Je considère cependant que la prouesse de se maintenir en positif dans un domaine aussi risqué que les paris sportifs a été relevée.

Développer des robots pour l'analyse et le placement automatique d'ordres financiers sans perte d'argent reste une véritable satisfaction et une première étape réussie.

14.2 Synthèse technique et organisationnelle

Sur le plan technique, l'utilisation d'une méthode agile de gestion du projet était indispensable car l'inverse m'aurait trop tôt enfermé dans des choix technologiques probablement trop limitant.

Dans un projet de ce type, constitué d'une grande partie de recherche et d'expérimentation, les choix faits ont assurés une adaptation progressive et un passage à l'échelle juste. Cela évite de perdre du temps dans la configuration d'environnements complexes et peut-être inadaptés avant d'avoir une meilleure idée de la solution finale.

Les compétences techniques dont je disposais au démarrage ont permis la mise en place de Sprint.

Étapes de chaque phase de Sprint :

- Phase d'idéation
- Phase de maturation
- Phase d'essai
- Phase de développement
- Phase de mise en production

Amélioration continue, à chaque nouveau Sprint :

- Les outils : CSV, puis MySQL, puis MySQL Cluster
- Les langages backend : Excel, puis Php, puis C++ sur certains batchs
- Les environnements : Local, serveur, cloud Rackspace
- Les interfaces fournisseurs : Crawl HTML, puis Rest API Betfair
- La supervision : Interface de pilotage, génération de logs, génération d'alerte, monitoring

Durant les 2 années de projet, cette aventure m'a permis d'approfondir encore mes connaissances techniques et de progresser significativement dans la mise au point de prototypes et leur passage au stade de produits de production. La capacité d'adaptation, pilotée par les nouveaux besoins rencontrés, aura été la clé de cette réussite technologique.

Il est intéressant de noter à quel point la méthode Agile s'applique naturellement sur ce type de projet.

Avec initialement un travail fortement orienté sur les analyses statistiques (équipes, points, période de l'année, nombre de buts, comparaison sur les années passées, ...), il m'a fallu évoluer rapidement vers de l'analyse financière en observant les évolutions des marchés, de leur volume et des côtes qui suffisaient en définitive à la détection des opportunités de gains.

Ce changement de direction a été accompagné d'un changement technologique, lié d'une part au développement de l'ensemble des nouveaux modules de connexion aux Webservices Betfair mais également aux évolutions du simulateur et aux nouveaux besoins beaucoup plus massif de stockage des informations collectées.

Je retiens comme étant le plus grand défi technique et intellectuel du projet, la difficulté à reproduire en production les résultats obtenus en simulation.

En effet, comme détaillé dans la section "Essais en simulateur", le travail sur des milliers de statistiques et des milliers de résultats sportifs a permis à de nombreuses reprises d'obtenir des prédictions spectaculaires, sans réussir à vérifier le comportement en production. C'est ce que nous appelions "faire chanter la base de données", car même sur une profondeur de plusieurs années, le travail d'optimisation algorithmique s'apparentait souvent à la découverte des meilleures formules pour exploiter les données existantes, et malheureusement pas nécessairement les nouvelles données en production. Ce fût l'un des constats les plus difficiles à admettre lorsque l'on traite des millions de données, et par conséquent le problème le plus difficile à contourner.

Je pense qu'il s'agit du même type de difficultés que celles rencontrées par les experts en météorologie dans leurs prévisions par apprentissage sur la base des données historiques.

14.3 Synthèse humaine

Une équipe commence à partir de deux personnes. Durant ces mois de projet, l'aventure de l'entrepreneuriat a été une forme de concrétisation du savoir-faire en relation humaine que nous savions apporter. Mon partenaire et moi-même avons pu mettre en oeuvre notre capacité à profiter des compétences de chacun tout en apportant un regard croisé indispensable à l'équilibre d'un projet et afin d'éviter toute dérive.

Savoir écouter, savoir communiquer, savoir proposer, savoir supporter sont autant de qualités dont il a fallu user dans ce contexte particulièrement éprouvant et même parfois stressant que sont les paris automatisés "haute-fréquence".

J'ai donc particulièrement apprécié ces moments de travail collaboratif sur les phases d'analyse des probabilités et d'études financières, tout en conservant la charge principale dans la réalisation et le développement du système Qwin d'un point de vue technique.

La phase de recul depuis la fin du projet et le fait de disposer d'un système complet, stable et toujours utilisable me laissent à penser que l'aventure n'en est peut-être qu'à ses débuts et que de nouvelles expérimentations pourraient voir le jour à partir de nouveaux jeux de données encore plus représentatifs dans la durée.

Chapitre 15

Annexes

15.1 Grandes phases de l'approche fonctionnelle du projet

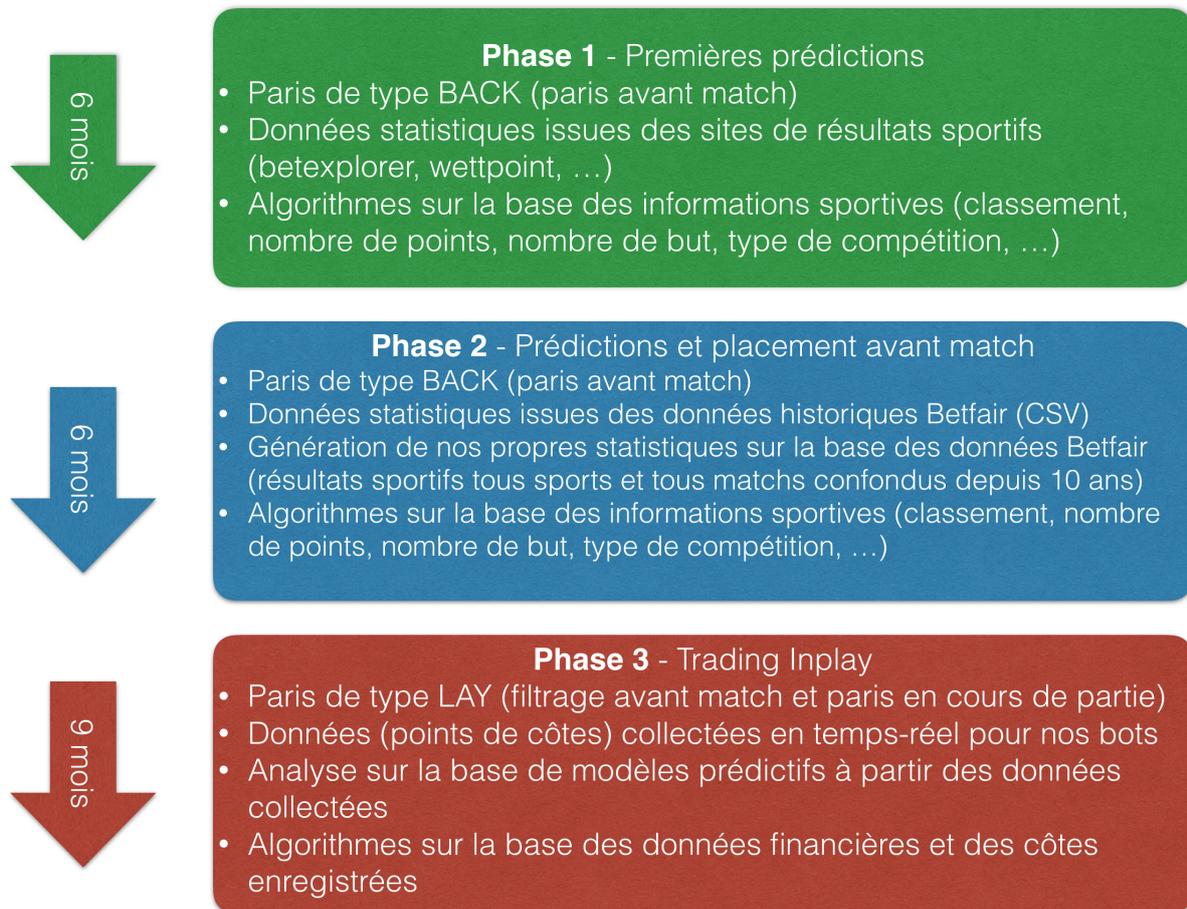


FIGURE 15.1 – Qwin, les phases fonctionnelles du projet

15.2 Exemples de placement de pari sur l'API Betfair

PLACE ORDER

CNX OK

```
object(stdClass)#4 (3) {
  ["status"]=> string(7) "SUCCESS" ["marketId"]=> string(11) "1.110165056"
  ["instructionReports"]=> array(1) {
    [0]=> object(stdClass)#5 (6) {
      ["status"]=> string(7) "SUCCESS" ["instruction"]=> object(stdClass)#6 (5) {
        ["orderType"]=> string(5) "LIMIT" ["selectionId"]=> int(3740233) ["handicap"]=>
          float(0) ["side"]=> string(4) "BACK" ["limitOrder"]=> object(stdClass)#7 (3) {
            ["size"]=> float(2) ["price"]=> float(1.83) ["persistenceType"]=> string(5) "LAPSE"
          } } ["betId"]=> string(11) "28410713506" ["placedDate"]=> string(24)
            "2013-07-22T20:36:51.000Z" ["averagePriceMatched"]=> float(0) ["sizeMatched"]=>
              float(0) } } }
```

Grâce à l'utilisation de la classe `bfxchange` développée dans le cadre du projet, le webservice Betfair retourne sa réponse au format JSON suite à un ordre de pari.

Cet exemple de placement de pari permet de comprendre le fonctionnement en 2 temps. D'abord la demande de placement de pari (réponse ci-dessus) et la validation de ce dernier (réponse ci-dessous). La lecture du champ `sizeMatched`, qui est passé à 2, signifie que la demande de pari à 2 euros sur une côte de 1.83 en BACK a pu être réalisée. Dans le mémoire la notion de Buckets est expliquée, il s'agit des montants disponibles par pari et par côte. Dans notre exemple, le bucket pour la côte 1.83 en BACK sur le marché `marketId` 1.110165056 et sur le runner (`selectionId`) 3740233 disposait d'au moins 2 euros disponibles, ce qui a permis de passer notre pari (`betId`) au statut `MATCHED` (100% du montant voulu a pu être placé).

CHECK CURRENT BET

CNX OK

```
object(stdClass)#4 (2) {
  ["currentOrders"]=> array(1) {
    [0]=> object(stdClass)#5 (18) {
      ["\glssymbol{betId}"]=> string(11) "28410713506" ["\gsl{marketId}"]=> string(11)
        "1.110165056" ["\glssymbol{selectionId}"]=> int(3740233) ["handicap"]=> float(0)
      ["priceSize"]=> object(stdClass)#6 (2) {
        ["price"]=> float(1.83) ["size"]=> float(2) } ["bspLiability"]=> float(0) ["side"]=>
        string(4) "BACK" ["status"]=> string(18) "EXECUTION_COMPLETE"
      ["persistenceType"]=> string(5) "LAPSE" ["orderType"]=> string(5) "LIMIT"
      ["placedDate"]=> string(24) "2013-07-22T20:36:51.000Z" ["averagePriceMatched"]=>
        float(1.83) ["sizeMatched"]=> float(2) ["sizeRemaining"]=> float(0)
      ["sizeLapsed"]=> float(0) ["sizeCancelled"]=> float(0) ["sizeVoided"]=> float(0)
      ["regulatorCode"]=> string(38) "MALTA LOTTERIES AND GAMBLING AUTHORITY" } }
      ["moreAvailable"]=> bool(false) }
```

15.3 Extrait des logs du système Qwin

```

date          log
2013-10-24 17:16:00 getBFBalance
2013-10-24 17:16:00 OK storing realtimeamount
2013-10-24 17:16:00 entering updateStatusInDb
2013-10-24 17:16:00 OK nbCurrentBet still the same number of current bet in DB : 28
2013-10-24 17:16:00 placebet nbbet 28 / betNumber 30
2013-10-24 17:16:00 OK getRange for football
2013-10-24 17:16:00 OK getRange for lay-football
2013-10-24 17:16:00 OK - Match has to be checked for backing / time control : 1.111477401
2013-10-24 17:16:00 placebet Try to place bet with marketid '1.111477401'
2013-10-24 17:16:00 OK this->myRangeArray[pSport][0] properly set
2013-10-24 17:16:00 OK this->myRangeArray[pSport][1] properly set
2013-10-24 17:16:00 Controlling time
2013-10-24 17:16:00 It is Thu 24 18:16 and the kickoff is at Thu 24 17:00
2013-10-24 17:16:00 Is is too late for betting for marketid:1.111477401
2013-10-24 17:16:00 placebet -> it is not the time to bet for '1.111477401'
2013-10-24 17:16:00 OK - Match has to be checked for backing / time control : 1.111478982
2013-10-24 17:16:00 placebet Try to place bet with marketid '1.111478982'
2013-10-24 17:16:00 OK this->myRangeArray[pSport][0] properly set
2013-10-24 17:16:00 OK this->myRangeArray[pSport][1] properly set
2013-10-24 17:16:00 Controlling time
2013-10-24 17:16:00 It is Thu 24 18:16 and the kickoff is at Thu 24 19:05
2013-10-24 17:16:00 Is is too late for betting for marketid:1.111478982
2013-10-24 17:16:00 placebet -> it is not the time to bet for '1.111478982'
2013-10-24 17:16:00 KO - Market closed for marketid : 1.111483200
2013-10-24 17:16:00 OK - Match has to be checked for backing / time control : 1.111493464
2013-10-24 17:16:00 placebet Try to place bet with marketid '1.111493464'
2013-10-24 17:16:00 OK this->myRangeArray[pSport][0] properly set
2013-10-24 17:16:00 OK this->myRangeArray[pSport][1] properly set
2013-10-24 17:16:00 Controlling time
2013-10-24 17:16:00 It is Thu 24 18:16 and the kickoff is at Sat 26 13:00
2013-10-24 17:16:00 Flotation point check - winback:2.28,winlay:2.4,lossback:3.7 - winback x lossback:8.436
2013-10-24 17:16:00 Flotation point reached for marketid : 1.111493464
2013-10-24 17:16:00 It is time for betting but range control failed for marketid : 1.111493464
2013-10-24 17:16:00 placebet -> it is not the time to bet for '1.111493464'
2013-10-24 17:16:00 OK - Match has to be checked for backing / time control : 1.111493733
2013-10-24 17:16:00 placebet Try to place bet with marketid '1.111497368'
2013-10-24 17:16:00 OK this->myRangeArray[pSport][0] properly set
2013-10-24 17:16:00 OK this->myRangeArray[pSport][1] properly set
2013-10-24 17:16:00 Controlling time
2013-10-24 17:16:00 It is Thu 24 18:16 and the kickoff is at Sat 26 06:30
2013-10-24 17:16:00 Flotation point check - winback:1.05,winlay:0,lossback:1.05 - winback x lossback:1.1025
2013-10-24 17:16:00 Flotation point reached for marketid : 1.111497368
2013-10-24 17:16:00 It is time for betting but range control failed for marketid : 1.111497368
2013-10-24 17:16:00 placebet -> it is not the time to bet for '1.111497368'
2013-10-24 17:16:00 OK - Match has to be checked for backing / time control : 1.111517238
2013-10-24 17:16:00 placebet Try to place bet with marketid '1.111517238'
2013-10-24 17:16:00 OK this->myRangeArray[pSport][0] properly set
2013-10-24 17:16:00 OK this->myRangeArray[pSport][1] properly set
2013-10-24 17:16:00 Controlling time
2013-10-24 17:16:00 It is Thu 24 18:16 and the kickoff is at Fri 25 16:00
2013-10-24 17:16:00 Flotation point check - winback:1.06,winlay:0,lossback:1.06 - winback x lossback:1.1236
2013-10-24 17:16:00 Flotation point reached for marketid : 1.111517238
2013-10-24 17:16:00 It is time for betting but range control failed for marketid : 1.111517238
2013-10-24 17:16:00 placebet -> it is not the time to bet for '1.111517238'
2013-10-24 17:16:00 OK - Match has to be checked for backing / time control : 1.111556451
2013-10-24 17:16:00 placebet Try to place bet with marketid '1.111556451'
2013-10-24 17:16:00 OK this->myRangeArray[pSport][0] properly set
2013-10-24 17:16:00 OK this->myRangeArray[pSport][1] properly set
2013-10-24 17:16:00 Controlling time
2013-10-24 17:16:00 It is Thu 24 18:16 and the kickoff is at Sat 26 10:00
2013-10-24 17:16:00 Flotation point check - winback:1.05,winlay:500,lossback:1.05 - winback x lossback:1.1025
2013-10-24 17:16:00 Flotation point not reached for marketid : 1.111556451
2013-10-24 17:16:00 placebet -> it is not the time to bet for '1.111556451'
2013-10-24 17:16:00 OK placebet - BF Error : 0 - Nb placedbet : 0 - Nb bet for the session : 28
2013-10-24 17:16:00 API contacted with action=tradecontrol
2013-10-24 17:16:01 OK - Checking if match 1.111479541 (PAOK - Maccabi Haifa) has been traded
2013-10-24 17:16:01 OK - Match 1.111509732 not traded, maybe a new lay to place
2013-10-24 17:16:01 OK - Checking if match 1.111494346 (Istanbul BB - Kahramanmaraşspor) has been traded
2013-10-24 17:16:01 OK - Match 1.111494346 not traded, maybe a new lay to place
2013-10-24 17:16:01 OK - Checking if match 1.111497956 (FC Heidenheim - Wehen Wiesbaden) has been traded
2013-10-24 17:16:01 OK - Match 1.111497956 not traded, maybe a new lay to place
2013-10-24 17:16:01 OK - Checking if match 1.111502035 (Leverkusen - Augsburg) has been traded
2013-10-24 17:16:01 OK - Match 1.111502035 not traded, maybe a new lay to place
2013-10-24 17:16:01 OK - Checking if match 1.111477270 (Lyon - Rijeka) has been traded
2013-10-24 17:16:01 OK - Match Already traded - 1.111477270
2013-10-24 17:16:01 OK - Checking if match 1.111478517 (Anzhi Makhachkala - Tromso) has been traded
2013-10-24 17:16:01 OK - Match 1.111478517 not traded, maybe a new lay to place

```

2013-10-24 17:16:01 OK - Checking if match 1.111477532 (Swansea - Kuban Krasnodar) has been traded
2013-10-24 17:16:01 OK - Match 1.111477008 not traded, maybe a new lay to place
2013-10-24 17:16:01 OK - Checking if match 1.111478610 (Apollon L. - Lazio) has been traded
2013-10-24 17:16:01 OK - Match 1.111478610 not traded, maybe a new lay to place
2013-10-24 17:16:01 OK - Checking if match 1.111512451 (Isidro Metapan - Los Angeles) has been traded
2013-10-24 17:16:01 OK - Match 1.111512451 not traded, maybe a new lay to place
2013-10-24 17:16:01 OK - Checking if match 1.111569930 (FC Mertskhali Ozurgeti - FC Dinamo Tbilisi II) has been traded
2013-10-24 17:16:01 OK - Match 1.111569930 not traded, maybe a new lay to place
2013-10-24 17:16:01 OK - Checking if match 1.111509369 (Gent - Cercle Brugge) has been traded
2013-10-24 17:16:01 OK - Match 1.111509369 not traded, maybe a new lay to place
2013-10-24 17:16:01 OK - Checking if match 1.111515508 (FC Vaslui - Uni Cluj) has been traded
2013-10-24 17:16:01 OK - Match 1.111515508 not traded, maybe a new lay to place
2013-10-24 17:16:01 OK - Checking if match 1.111508582 (Braga - Academica) has been traded
2013-10-24 17:16:01 OK - Match 1.111508582 not traded, maybe a new lay to place
2013-10-24 17:16:01 OK - Checking if match 1.111491119 (Debrecen - Lombard Papa) has been traded
2013-10-24 17:16:01 OK - Match 1.111491119 not traded, maybe a new lay to place
2013-10-24 17:16:01 OK - Checking if match 1.111517441 (Afan Lido - Carmarthen Town) has been traded
2013-10-24 17:16:01 TradeControl :: it is not the time to lay for marketid:'1.111491119' - currentlay:1.35 - goldenratelay:2.3
2013-10-24 17:16:01 OK - Match has to be checked for trading control : 1.111499968
2013-10-24 17:16:01 OK - Testing marketid:'1.111499968' - Current Back:1.56 - Current Lay:1.57 - Magic:1.42 - Golden:3.12
2013-10-24 17:16:01 TradeControl :: it is not the time to lay for marketid:'1.111499968' - currentlay:1.57 - goldenratelay:3.32
2013-10-24 17:16:01 OK - Match has to be checked for trading control : 1.111507344
2013-10-24 17:16:01 OK - Testing marketid:'1.111507344' - Current Back:1.39 - Current Lay:1.4 - Magic:1.31 - Golden:2.78
2013-10-24 17:16:01 TradeControl :: it is not the time to lay for marketid:'1.111507344' - currentlay:1.4 - goldenratelay:2.98
2013-10-24 17:16:01 OK - Match has to be checked for trading control : 1.111517441
2013-10-24 17:16:01 OK - Testing marketid:'1.111517441' - Current Back:1.58 - Current Lay:1.87 - Magic:1.42 - Golden:3.14
2013-10-24 17:16:01 TradeControl :: it is not the time to lay for marketid:'1.111517441' - currentlay:1.87 - goldenratelay:3.34
2013-10-24 17:16:01 OK - Match has to be checked for trading control : 1.111517447
2013-10-24 17:16:01 OK - Testing marketid:'1.111517447' - Current Back:1.21 - Current Lay:1.31 - Magic:1.2 - Golden:2.42
2013-10-24 17:16:01 TradeControl :: it is not the time to lay for marketid:'1.111517447' - currentlay:1.31 - goldenratelay:2.62
2013-10-24 17:16:01 OK - Match has to be checked for trading control : 1.111561441
2013-10-24 17:16:01 OK - Testing marketid:'1.111561441' - Current Back:1.29 - Current Lay:1.59 - Magic:1.25 - Golden:2.58
2013-10-24 17:16:01 TradeControl :: it is not the time to lay for marketid:'1.111561441' - currentlay:1.59 - goldenratelay:2.78
2013-10-24 17:16:02 API contacted with action=getqwinbotvalue
2013-10-24 17:16:03 API contacted with action=getqwinbotvalue
2013-10-24 17:16:04 API contacted with action=getqwinbotvalue
2013-10-24 17:16:04 API contacted with action=marketpoints
2013-10-24 17:16:04 OK - Entering marketPoints
2013-10-24 17:16:04 OK - Retrieving marketPoints for matchs to bet or already bet
2013-10-24 17:16:05 API contacted with action=tradecontrol
2013-10-24 17:16:05 OK - Checking if match 1.111479541 (PAOK - Maccabi Haifa) has been traded
2013-10-24 17:16:05 OK - Match 1.111479541 not traded, maybe a new lay to place
2013-10-24 17:16:05 OK - Checking if match 1.111479635 (Red Bull Salzburg - Standard) has been traded
2013-10-24 17:16:05 OK - Match 1.111479635 not traded, maybe a new lay to place
2013-10-24 17:16:05 OK - Checking if match 1.111516349 (Fenerbahce - Gaziantepspor) has been traded
2013-10-24 17:16:05 OK - Match 1.111516349 not traded, maybe a new lay to place

15.4 Trading Inplay, apprentissage et modèles prédictifs

Dans le mémoire ont été détaillées les différentes méthodes de paris, les étapes du projet et l'utilisation des statistiques pour le filtrage des matchs avant le coup d'envoi.

Le trading inplay a souvent été mentionné et représente concrètement le mode de pari le plus performant.

Il nécessite une prise en compte des données temporelles (analyse des périodes les plus intéressantes pour le placement de pari au cours d'un match par apprentissage sur l'évolution des côtes en cours de partie), des données statistiques (performances des équipes, à l'extérieur ou à domicile) et enfin les données financières (montants disponibles et activités au niveau des buckets).

C'est par l'ensemble des phases d'apprentissage et d'exécution des programmes de détection des modèles de prédiction (compréhension de l'évolution d'un marché, influence du nombre de parieurs, influence des périodes d'un match, influence des montants en jeu, ...) que les algorithmes ont pu être définis et progressivement affinés.

Détail ci-après de mes conclusions obtenues des nombreuses analyses statistiques et probabilistes dans le cadre du projet. Ce travail m'a permis de dégager les règles fondamentales permettant d'obtenir des résultats dans le milieu des marchés sportifs.

FILTRAGE AVANT MATCH ABDIFF

- Peu d'impact si à domicile ou à l'extérieur (notion inexistante au tennis)
- Calcul d'écart entre les côtes de BACK du runner A et du runner B (exemple ABDIFF > 3)
- Pondération en fonction des montants en jeu

STRATEGIE SUR UN EXEMPLE DE MATCH DE FOOTBALL MODE INPLAY

- Période 15 minutes avant ou après la mi-temps
- Pari uniquement en LAY sur favori si ce dernier perd avec un but d'écart (mesure en fonction de l'écart des côtes de BACK)
- Côte en LAY maximum à 3 (au delà le risque devient trop grand)

STRATEGIE SUR UN EXEMPLE DE MATCH DE TENNIS PARTIE MODE INPLAY

- Tout au long de la partie
- Pari uniquement en LAY sur favori si ce dernier a une côte supérieure à 2 (il vient de perdre un jeu)
- Côte en LAY maximum à 3 (au delà le risque devient trop grand)

MATCHED ONLY

- Chercher en permanence l'ensemble des buckets disponibles
- Ne placer que des paris qui deviendront MATCHED à 100%
- Répartir la mise sur l'ensemble des buckets disponibles en dessous de la côte choisie en LAY

DISPERSION

- Préférer le volume à un taux de réussite de 90%
- Diversifier sur l'ensemble des championnats et compétitions
- Ne pas se limiter géographiquement
- Réduire le montant mais multiplier les paris en inplay

HAUTE FREQUENCE

- Assurer une proximité entre votre serveur et l'Exchange
- Analyser les côtes le plus fréquemment possible en cours de match (min 1/minute)

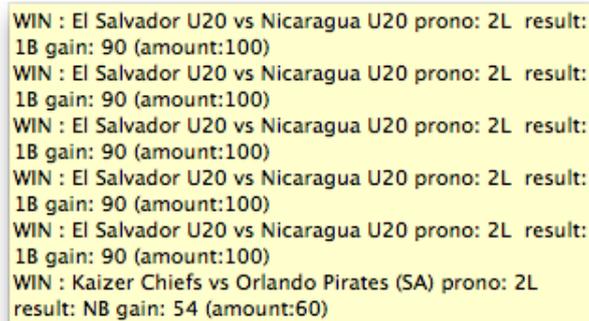
15.5 Rewind, exemple de résultats en sortie du simulateur

Le simulateur permettait une grande flexibilité de configuration.

En fonction de la période de temps et des algorithmes en phase de tests, la durée d'une exécution Rewind variait fortement, pouvant prendre quelques minutes comme plusieurs heures.

Chaque journée de calcul traitée par Rewind disposait du détail des paris placés comme on le remarque sur la capture suivante.

On y découvre l'ensemble des paris de la journée, leur type et les montants associés en fonction des résultats.



```
WIN : El Salvador U20 vs Nicaragua U20 pronos: 2L result:  
1B gain: 90 (amount:100)  
WIN : El Salvador U20 vs Nicaragua U20 pronos: 2L result:  
1B gain: 90 (amount:100)  
WIN : El Salvador U20 vs Nicaragua U20 pronos: 2L result:  
1B gain: 90 (amount:100)  
WIN : El Salvador U20 vs Nicaragua U20 pronos: 2L result:  
1B gain: 90 (amount:100)  
WIN : El Salvador U20 vs Nicaragua U20 pronos: 2L result:  
1B gain: 90 (amount:100)  
WIN : Kaizer Chiefs vs Orlando Pirates (SA) pronos: 2L  
result: NB gain: 54 (amount:60)
```

FIGURE 15.2 – Rewind, détail des paris pour chaque journée

C'est pour chacune des cellules du tableau de sortie Rewind que le détail des opérations apparaissait.

Exemple de résultats en sortie du simulateur

// QWin Rewind Back Only Fish Tool Tool .. the graal quest ...
 ^^^^^^^^^^^ Teiss&Fredoun @2014 ^^^^^^^^^^^

Options :
 preghost=[0|1] : Disable or Enable Ghosting mode for pre kickoff bet (default 1)
 rank=[0|1] : Disable or Enable rank table usage (default 1)
 ff=[0|1] : Disable or Enable Follow Forecast usage (default 1)
 notinplay=[0|1] : Disable or Enable not inplay matches (default 1)
 memory=[0|1] : Disable or Enable splitting bet throughout snapshots usage (default 1)
 simulay=[0|1] : Disable or Enable simulay for matchs before the 22/09/2014 (default 0)
 from=YYYY-mm-dd&to=YYYY-mm-dd : set a period to rewind
 nbalgo=[0-x] : Indicate the number of algo to rewind
 fjump=[1-x] : Indicate interval between 2 FF try (1 min minimum and default)
 mtype=[ODDS|DNB|ALL] : Specify the type of market you want (default ODDS)
 sport=[football|tennis|all] : Specify the type of market you want (default all)
 marketid=x.xxxxxx : Force rewind for one marketid (default all matchs)
 player=abcd : Force rewind for one player (default all players)
 debug=[0|1] : Disable or Enable debug mode (default 0)

==> Request rewind from 2014-09-23 (1411430400) to 2014-12-14 (1418515200)

nb algo : 2 - betamount : 30

```
select * from qwinrank where weeknb = week(from_unixtime(1411430400), 1) and year= year(from_unixtime(1411430400))
select * from qwinrank where weeknb = week(from_unixtime(1411430400), 1) and year= year(from_unixtime(1411430400)) already done :)
for week1411430400 select * from qwinrank where weeknb = week(from_unixtime(1411948800), 1) and year=
year(from_unixtime(1411948800))
select * from qwinrank where weeknb = week(from_unixtime(1411948800), 1) and year= year(from_unixtime(1411948800)) already done :)
for week1411948800 select * from qwinrank where weeknb = week(from_unixtime(1412553600), 1) and year=
year(from_unixtime(1412553600))
select * from qwinrank where weeknb = week(from_unixtime(1412553600), 1) and year= year(from_unixtime(1412553600)) already done :)
for week1412553600 select * from qwinrank where weeknb = week(from_unixtime(1413158400), 1) and year=
year(from_unixtime(1413158400))
select * from qwinrank where weeknb = week(from_unixtime(1413158400), 1) and year= year(from_unixtime(1413158400)) already done :)
for week1413158400 select * from qwinrank where weeknb = week(from_unixtime(1413763200), 1) and year=
year(from_unixtime(1413763200))
select * from qwinrank where weeknb = week(from_unixtime(1413763200), 1) and year= year(from_unixtime(1413763200)) already done :)
for week1413763200 select * from qwinrank where weeknb = week(from_unixtime(1414368000), 1) and year=
year(from_unixtime(1414368000))
select * from qwinrank where weeknb = week(from_unixtime(1414368000), 1) and year= year(from_unixtime(1414368000)) already done :)
for week1414368000 select * from qwinrank where weeknb = week(from_unixtime(1414972800), 1) and year=
year(from_unixtime(1414972800))
select * from qwinrank where weeknb = week(from_unixtime(1414972800), 1) and year= year(from_unixtime(1414972800)) already done :)
for week1414972800 select * from qwinrank where weeknb = week(from_unixtime(1415577600), 1) and year=
year(from_unixtime(1415577600))
select * from qwinrank where weeknb = week(from_unixtime(1415577600), 1) and year= year(from_unixtime(1415577600)) already done :)
for week1415577600 select * from qwinrank where weeknb = week(from_unixtime(1416182400), 1) and year=
year(from_unixtime(1416182400))
select * from qwinrank where weeknb = week(from_unixtime(1416182400), 1) and year= year(from_unixtime(1416182400)) already done :)
for week1416182400 select * from qwinrank where weeknb = week(from_unixtime(1416787200), 1) and year=
year(from_unixtime(1416787200))
select * from qwinrank where weeknb = week(from_unixtime(1416787200), 1) and year= year(from_unixtime(1416787200)) already done :)
for week1416787200 select * from qwinrank where weeknb = week(from_unixtime(1417392000), 1) and year=
year(from_unixtime(1417392000))
select * from qwinrank where weeknb = week(from_unixtime(1417392000), 1) and year= year(from_unixtime(1417392000)) already done :)
for week1417392000 select * from qwinrank where weeknb = week(from_unixtime(1417996800), 1) and year=
year(from_unixtime(1417996800))
select * from qwinrank where weeknb = week(from_unixtime(1417996800), 1) and year= year(from_unixtime(1417996800)) already done :)
for week1417996800
```

Date	Algo GhostFFKiller	Algo GhostFFDegun
09-23-2014	nbbet:12 \$:56.1 (dexp:503)	nbbet:65 \$:75.36 (dexp:714)
09-24-2014	nbbet:30 \$:340.34 (dexp:2518)	nbbet:87 \$:479.28 (dexp:1079)
09-25-2014	nbbet:16 \$:405.79 (dexp:569)	nbbet:43 \$:550.34 (dexp:171)
09-26-2014	nbbet:9 \$:405.79 (dexp:0)	nbbet:75 \$:421.89 (dexp:874)
09-27-2014	nbbet:34 \$:499.76 (dexp:827)	nbbet:235 \$:463.49 (dexp:2429)
09-28-2014	nbbet:37 \$:591.39 (dexp:837)	nbbet:232 \$:928.06 (dexp:2705)
09-29-2014	nbbet:14 \$:654.97 (dexp:614)	nbbet:71 \$:834.44 (dexp:353)
09-30-2014	nbbet:12 \$:716.68 (dexp:534)	nbbet:49 \$:1075.67 (dexp:669)

10-01-2014	nbbet:13 \$:697.39 (dexp:668)	nbbet:70 \$:1485.2 (dexp:1238)
10-02-2014	nbbet:10 \$:697.39 (dexp:0)	nbbet:46 \$:1819.04 (dexp:1056)
10-03-2014	nbbet:21 \$:763.77 (dexp:574)	nbbet:74 \$:1654.14 (dexp:1075)
10-04-2014	nbbet:46 \$:874.1 (dexp:924)	nbbet:271 \$:1945.86 (dexp:817)
10-05-2014	nbbet:40 \$:1077 (dexp:1761)	nbbet:232 \$:2569.4 (dexp:2231)
10-06-2014	nbbet:9 \$:1077 (dexp:0)	nbbet:47 \$:2730.33 (dexp:609)
10-07-2014	nbbet:7 \$:1103.18 (dexp:226)	nbbet:50 \$:2858.43 (dexp:335)
10-08-2014	nbbet:10 \$:1065.06 (dexp:555)	nbbet:35 \$:2686.36 (dexp:517)
10-09-2014	nbbet:10 \$:1121.16 (dexp:526)	nbbet:20 \$:2686.36 (dexp:0)
10-10-2014	nbbet:17 \$:1209.99 (dexp:784)	nbbet:37 \$:2789.21 (dexp:263)
10-11-2014	nbbet:22 \$:1220.27 (dexp:105)	nbbet:209 \$:2325.87 (dexp:2591)
10-12-2014	nbbet:18 \$:1320.32 (dexp:949)	nbbet:135 \$:1961.43 (dexp:1692)
10-13-2014	nbbet:5 \$:1320.32 (dexp:0)	nbbet:17 \$:2088.59 (dexp:308)
10-14-2014	nbbet:10 \$:1376.42 (dexp:543)	nbbet:17 \$:2239.12 (dexp:463)
10-15-2014	nbbet:7 \$:1376.42 (dexp:0)	nbbet:37 \$:2467.26 (dexp:618)
10-16-2014	nbbet:5 \$:1376.42 (dexp:0)	nbbet:18 \$:2578.53 (dexp:301)
10-17-2014	nbbet:11 \$:1396.05 (dexp:189)	nbbet:78 \$:2259.69 (dexp:1740)
10-18-2014	nbbet:45 \$:1542.85 (dexp:1336)	nbbet:292 \$:3176.95 (dexp:2974)
10-19-2014	nbbet:35 \$:1552.2 (dexp:78)	nbbet:248 \$:3534.12 (dexp:1004)
10-20-2014	nbbet:8 \$:1552.2 (dexp:0)	nbbet:57 \$:3617.33 (dexp:309)
10-21-2014	nbbet:6 \$:1603.62 (dexp:423)	nbbet:53 \$:3935.23 (dexp:744)
10-22-2014	nbbet:16 \$:1863.55 (dexp:2255)	nbbet:63 \$:3593.66 (dexp:1252)
10-23-2014	nbbet:10 \$:1972.95 (dexp:982)	nbbet:37 \$:3692.88 (dexp:3389)
10-24-2014	nbbet:15 \$:2001 (dexp:243)	nbbet:96 \$:3333.57 (dexp:3238)
10-25-2014	nbbet:44 \$:2215.11 (dexp:1890)	nbbet:250 \$:4323.54 (dexp:3935)
10-26-2014	nbbet:56 \$:2702.25 (dexp:4516)	nbbet:227 \$:5867.55 (dexp:5892)
10-27-2014	nbbet:10 \$:2780.79 (dexp:699)	nbbet:52 \$:5880.12 (dexp:1529)
10-28-2014	nbbet:18 \$:2780.79 (dexp:0)	nbbet:56 \$:5904.06 (dexp:256)
10-29-2014	nbbet:38 \$:3102.43 (dexp:2970)	nbbet:68 \$:6104.15 (dexp:626)
10-30-2014	nbbet:11 \$:3325.89 (dexp:2066)	nbbet:28 \$:6307.98 (dexp:704)
10-31-2014	nbbet:12 \$:3366.1 (dexp:409)	nbbet:80 \$:7021.55 (dexp:2417)
11-01-2014	nbbet:44 \$:3755.06 (dexp:3591)	nbbet:235 \$:6424.8 (dexp:2821)
11-02-2014	nbbet:31 \$:3846.45 (dexp:909)	nbbet:258 \$:7016.36 (dexp:6328)
11-03-2014	nbbet:9 \$:4007.27 (dexp:1564)	nbbet:52 \$:7021.61 (dexp:3242)
11-04-2014	nbbet:8 \$:4007.27 (dexp:0)	nbbet:51 \$:7542.73 (dexp:1483)
11-05-2014	nbbet:14 \$:4007.27 (dexp:0)	nbbet:36 \$:6515.22 (dexp:3367)

11-06-2014	nbbet:12 \$:4206.43 (dexp:1792)	nbbet:34 \$:6526.44 (dexp:31)
11-07-2014	nbbet:16 \$:4269.07 (dexp:553)	nbbet:70 \$:7306.15 (dexp:2283)
11-08-2014	nbbet:42 \$:4292.03 (dexp:873)	nbbet:231 \$:7869.28 (dexp:4624)
11-09-2014	nbbet:43 \$:4466.87 (dexp:1540)	nbbet:218 \$:7819.72 (dexp:4424)
11-10-2014	nbbet:6 \$:4466.87 (dexp:0)	nbbet:11 \$:7850.57 (dexp:72)
11-11-2014	nbbet:7 \$:4466.87 (dexp:0)	nbbet:32 \$:7642.07 (dexp:209)
11-12-2014	nbbet:11 \$:3924.3 (dexp:726)	nbbet:23 \$:7797.28 (dexp:430)
11-13-2014	nbbet:10 \$:4006.58 (dexp:822)	nbbet:20 \$:7797.28 (dexp:0)
11-14-2014	nbbet:11 \$:4006.58 (dexp:0)	nbbet:50 \$:8383.53 (dexp:1555)
11-15-2014	nbbet:21 \$:4339.44 (dexp:3022)	nbbet:183 \$:8674.66 (dexp:3523)
11-16-2014	nbbet:18 \$:4618.07 (dexp:2489)	nbbet:158 \$:8088.19 (dexp:4419)
11-17-2014	nbbet:6 \$:4718.35 (dexp:960)	nbbet:33 \$:8374.3 (dexp:873)
11-18-2014	nbbet:9 \$:4806.24 (dexp:798)	nbbet:57 \$:8910.99 (dexp:1655)
11-19-2014	nbbet:10 \$:4824.94 (dexp:152)	nbbet:31 \$:8809.98 (dexp:2383)
11-20-2014	nbbet:5 \$:4824.94 (dexp:0)	nbbet:23 \$:9208.29 (dexp:1237)
11-21-2014	nbbet:6 \$:4968.93 (dexp:1281)	nbbet:73 \$:8501.34 (dexp:1492)
11-22-2014	nbbet:41 \$:5195.9 (dexp:2111)	nbbet:245 \$:9216.79 (dexp:5231)
11-23-2014	nbbet:23 \$:4972.39 (dexp:1147)	nbbet:217 \$:9117.4 (dexp:1764)
11-24-2014	nbbet:9 \$:5188.37 (dexp:2036)	nbbet:35 \$:9311.88 (dexp:587)
11-25-2014	nbbet:4 \$:5188.37 (dexp:0)	nbbet:45 \$:9571.81 (dexp:607)
11-26-2014	nbbet:4 \$:5188.37 (dexp:0)	nbbet:30 \$:9333.56 (dexp:238)
11-27-2014	nbbet:7 \$:5203.33 (dexp:126)	nbbet:18 \$:9528.04 (dexp:625)
11-28-2014	nbbet:8 \$:4999.42 (dexp:311)	nbbet:72 \$:10208.39 (dexp:3615)
11-29-2014	nbbet:40 \$:5058.32 (dexp:551)	nbbet:245 \$:11424.82 (dexp:3778)
11-30-2014	nbbet:23 \$:5153.69 (dexp:839)	nbbet:174 \$:9670.65 (dexp:2768)
12-01-2014	nbbet:10 \$:5250 (dexp:920)	nbbet:29 \$:9760.41 (dexp:270)
12-02-2014	nbbet:9 \$:5265.89 (dexp:144)	nbbet:42 \$:10295.17 (dexp:2093)
12-03-2014	nbbet:16 \$:4320.39 (dexp:1994)	nbbet:50 \$:10358.75 (dexp:143)
12-04-2014	nbbet:6 \$:4320.39 (dexp:0)	nbbet:17 \$:10474.69 (dexp:451)
12-05-2014	nbbet:7 \$:4320.39 (dexp:0)	nbbet:59 \$:10584.82 (dexp:691)
12-06-2014	nbbet:40 \$:4511.2 (dexp:3318)	nbbet:179 \$:10826.84 (dexp:4600)
12-07-2014	nbbet:35 \$:4710.35 (dexp:1761)	nbbet:161 \$:11241.35 (dexp:1478)
12-08-2014	nbbet:6 \$:4710.35 (dexp:0)	nbbet:54 \$:11766.82 (dexp:1826)
12-09-2014	nbbet:10 \$:4761.78 (dexp:453)	nbbet:36 \$:11599.62 (dexp:943)
12-10-2014	nbbet:9 \$:4871.17 (dexp:986)	nbbet:23 \$:11599.62 (dexp:0)
12-11-2014	nbbet:1 \$:4871.17 (dexp:0)	nbbet:4 \$:11599.62 (dexp:0)

||12-12-2014| nbbet:0 \$:4871.17 (dexp:0) | nbbet:0 \$:11599.62 (dexp:0) ||

GhostFFKiller

==> The global balance : 4871.17 and you bet 1396 times - Wrong 122 times - Draw 0 times - No Money 1036 times

- Prediction : 91%

FF win/placed : 815/846

FF Bet no money : 1230

NitroEffect : 0

NitroEffectLoss : 0

[Single Graph](#)

GhostFFDegun

==> The global balance : 11599.62 and you bet 7471 times - Wrong 1795 times - Draw 0 times - No Money 1036 times

- Prediction : 76%

FF win/placed : 3388/4242

FF Bet no money : 1

NitroEffect : 0

NitroEffectLoss : 0

[Single Graph](#)

NbStat : 14683 - MaxLoss : 440490

15.6 Memento des commandes de configuration OpenVPN

```

yum install traceroute && yum install openvpn -y && yum install openvpn easy-rsa
mkdir -p /etc/openvpn/easy-rsa/keys; cp -rf /usr/share/openvpn/easy-rsa/2.0/*
    /etc/openvpn/easy-rsa

mkdir /etc/openvpn/keys; vi /etc/openvpn/my-vpn.conf
touch /etc/openvpn/ipp.txt
chown -R openvpn:openvpn /etc/openvpn
chmod 750 /etc/openvpn; chmod 750 /etc/openvpn/keys; chmod 640
    /etc/openvpn/*.{conf,txt}
chmod 640 /etc/openvpn/keys/*.{crt,pem}; chmod 600 /etc/openvpn/keys/*.key
restorecon -v -F -R /etc/openvpn && restorecon -v -F /var/log/openvpn*.log
ln -s /lib/systemd/system/openvpn@.service
    /etc/systemd/system/multi-user.target.wants/openvpn@my-vpn.service

cd /etc/openvpn/keys; openssl dhparam -out dh2048.pem 2048; openvpn --genkey --secret
    ta.key

cd /usr/share/easy-rsa/2.0/ && . ./vars && ./clean-all && ./build-ca &&
    ./build-key-server server && ./build-key macfred && ./build-key macju &&
    ./build-key qwinsrv
cd keys/; cp * /etc/openvpn/keys/

[root@qwinvpn openvpn]# cat my-vpn.conf
port 1194
proto udp
dev tun
ca /etc/openvpn/keys/ca.crt
cert /etc/openvpn/keys/server.crt
key /etc/openvpn/keys/server.key # This file should be kept secret
dh /etc/openvpn/keys/dh2048.pem
server 10.0.3.0 255.255.255.0
ifconfig-pool-persist ipp.txt
push "dhcp-option DOMAIN qwin.org"
push "dhcp-option SEARCH qwin.org"
keepalive 10 120
# tls-auth /etc/openvpn/keys/ta.key 0
comp-lzo
user openvpn
group openvpn
persist-key
persist-tun
status /var/log/openvpn-status.log
log-append /var/log/openvpn.log
verb 4
mute 20

```

Glossaire

ABDIFF Paramètre clé du système de prédiction. Il représente l'écart entre la côte de BACK du runner A et la côte de BACK du runner B permettant de mesurer la différence entre eux et les probabilité de victoire de l'un par rapport à l'autre. 70

API Service applicatif mettant à disposition des méthodes et des données à l'ensemble du système. 1, 21, 27, 29

BACK Mode de pari sur la probabilité de victoire d'un runner par rapport à l'autre (joueur ou équipe). 12, 37, 38

BETID Identifiant unique Betfair pour chaque pari placé. 67

BIG PICTURE Architecture logicielle haut-niveau du projet. 20

BUCKETS Ensemble des montants disponibles pour chaque côte d'un évènement. 38, 67

CRAWL Opération de téléchargement de contenus Web à partir d'une liste d'URL. 11

CSV Format de séparation des données en colonne dans un fichier texte grâce à un caractère spécial comme le point virgule. Ex : nom;prénom;âge. 12, 29

EXCHANGE Terme désignant une place de marché sportive. Equivalent au Stock Exchange dans le milieu boursier.. 1, 19, 23, 31, 33, 38, 71

GIT Utilitaire de gestion de versions décentralisé. 22

HHVM Moteur d'exécution PHP. 27

HMA Utilitaire de connexion VPN depuis une machine de développement. 24

INPLAY Mode de pari en cours de partie. 29, 34, 35

LAY Mode de pari sur la probabilité de la défaite ou de l'égalité d'un runner par rapport à l'autre (joueur ou équipe). 12, 37

MARKETID Identifiant unique Betfair pour chaque marché disponible d'un évènement sportif. Cela représente une possibilité de pari en BACK ou en LAY. 67

NOSQL Not Only SQL signifie le nouveau paradigme clé/valeur des systèmes de gestion de base de données distribuées. 9

PHP Langage de script à exécution côté serveur. 25

PREDIDAY Site de revente de pronostics ayant permis le lancement de l'activité et l'autofinancement pour partie. 29

QWIN Nom du système développé dans le cadre du projet dans sa globalité. 1, 28

REWIND Nom donné au simulateur du système Qwin. 27, 34, 48, 49, 58

RUNNER Terme utilisé pour désigner un joueur ou une équipe durant une partie. 67, 70

SELECTIONID Identifiant unique Betfair pour chaque runner (joueur au tennis ou équipe au football). 67

SIZEMATCHED Montant réellement placé sur un ordre de pari en fonction des sommes disponibles sur les buckets. 67

VM Machine virtuelle exécutant un système d'exploitation sur une machine hôte. 23

VPN Tunnel sécurisé entre deux réseaux. 8, 21, 57

Table des figures

5.1	Prediday, Portail de vente de pronostics	15
5.2	Prediday, Achats de pronostic via Paypal	16
5.3	Prediday, Grilles disponibles en fonction des probabilités des pronostics	17
5.4	Qwin, Big Picture	20
6.1	OSX, le logo	22
6.2	Git, le logo	22
6.3	Rackspace, le logo	23
6.4	Red Hat, le logo	23
6.5	HideMyAss, le logo	24
6.6	High-Level-Design ; vue globale de l'architecture réseau	24
6.7	Logo PHP	25
6.8	Logo C++	25
6.9	MySQL, le logo	25
6.10	Fonctionnement de l'auto-sharding MySQL	26
6.11	Apache, le logo	26
6.12	HHVM, le logo	27
7.1	Betfair, Exemple de répartition par Bucket d'un évènement	39
7.2	BetTrader, Exemple de suivi de l'évolution des côtes et des Buckets	39
8.1	Betfair, portail des développeurs	43
8.2	Qwinboard, tableau de bord et administration	45
8.3	Qwinboard, affichage depuis un téléphone mobile	46
8.4	Qwinboard, tableau détaillé des paris	46
8.5	Rewind, génération de graphique sur l'évolution de la balance financière	49
8.6	QwinLib, librairies transverses du système Qwin	50
9.1	QwinTU, liste des scripts de test	54
10.1	Qwin log, outil d'investigation	55
12.1	Rewind Multithread, capture de l'utilisation de l'ensemble des coeurs du CPU	59
15.1	Qwin, les phases fonctionnelles du projet	66
15.2	Rewind, détail des paris pour chaque journée	72

Résumé

QWIN - Système automatisé de prédiction, trading et placement de paris sportifs

Loin d'être un jeu, le monde des paris sportifs est une véritable entreprise financière où les émotions laissent place à la technique, à l'apprentissage et à la prédiction. Après quelques expérimentations réussies, un travail d'étude et de recherche a été entrepris afin de découvrir les mécanismes des places de marchés sportifs. Les exigences initiales du projet précisaient clairement que le système Qwin devait être autonome et ne nécessiter aucune intervention humaine. L'objectif était simple, développer un système de prédiction et de placement de paris sportifs, basé sur une analyse des performances des équipes, et des évolutions financières des marchés en temps-réel. Durant 2 années, un travail de recherche, d'analyse, de conception, de développement et d'exploitation a été réalisé et a permis d'acquérir la maîtrise du fonctionnement des marchés sportifs ainsi que de mettre au point un ensemble de programmes pleinement fonctionnels, qualifiés, testés et performants autour de l'API de l'Exchange Betfair. Il a néanmoins fallu, par manque de fonds, mettre en standby le système, car, pour disposer des meilleures prédictions, une profondeur de base de données de plusieurs années est indispensable au simulateur. Le temps a manqué. Les bases utilisées au démarrage du projet ne contenaient pas assez d'information. Une phase de collecte supplémentaire de 2 ans aurait été nécessaire pour assurer une fiabilité des prédictions et une remise en route des applications. Malgré des résultats plus qu'encourageants, seulement 10 mois de données complètes ont pu être collectées. Le projet Qwin redémarrera très certainement. Un long travail ainsi qu'un investissement conséquent ont été nécessaires à la mise en place de ce projet. Ce mémoire décrit l'ensemble des phases du projet, détaille la partie développement et le fonctionnement des marchés sportifs et donne une appréciation sur ce qu'il faut retenir des méthodes et techniques d'apprentissage et de prédiction.

Mots clés :

pari, prédiction, bourse, robot, intelligence artificielle, megadonnées, service web, apprentissage.

QWIN - Automated forecast and trading system in sport exchanges

Far from being a game, betting exchange world is a real business where there is no place for something else than technical capabilities to learn and forecast. After some successful experiments, a study and research phase has been started in order to discover betting exchange mechanisms. The most important requirement was to build Qwin as a fully automated system, with no human interaction. Goal was simply to conceive a forecast and trading system based on player performance and real-time financial analysis. During 2 years, research, analyse, conception and development activities allowed to master how betting exchange works, using the Betfair API. To get the best possible predictions, a long time is needed in collecting data. Time ran out, and the system had to be stopped by lack of consistent data of financial resources. Initial set of information was enough to start projet but not to be confident in production despite encouraging results. 2 years collecting data was expected, and only 10 months were available. But no doubt the system will restart because it's fully operational and it has proved it provides high performance forecasts. The thesis is describing in depth all the project phases, how betting exchange works, how the conception and development has been realised, based on today computing technologies. It provides a detailed appreciation on what it's important to remember regarding methods for predicting : what, how and when.

Keywords :

betting, forecast, trading, automation, artificial intelligence, big data, webservice, deep learning.