Le métier de Data Scientist

Amel SAHLI, PhD

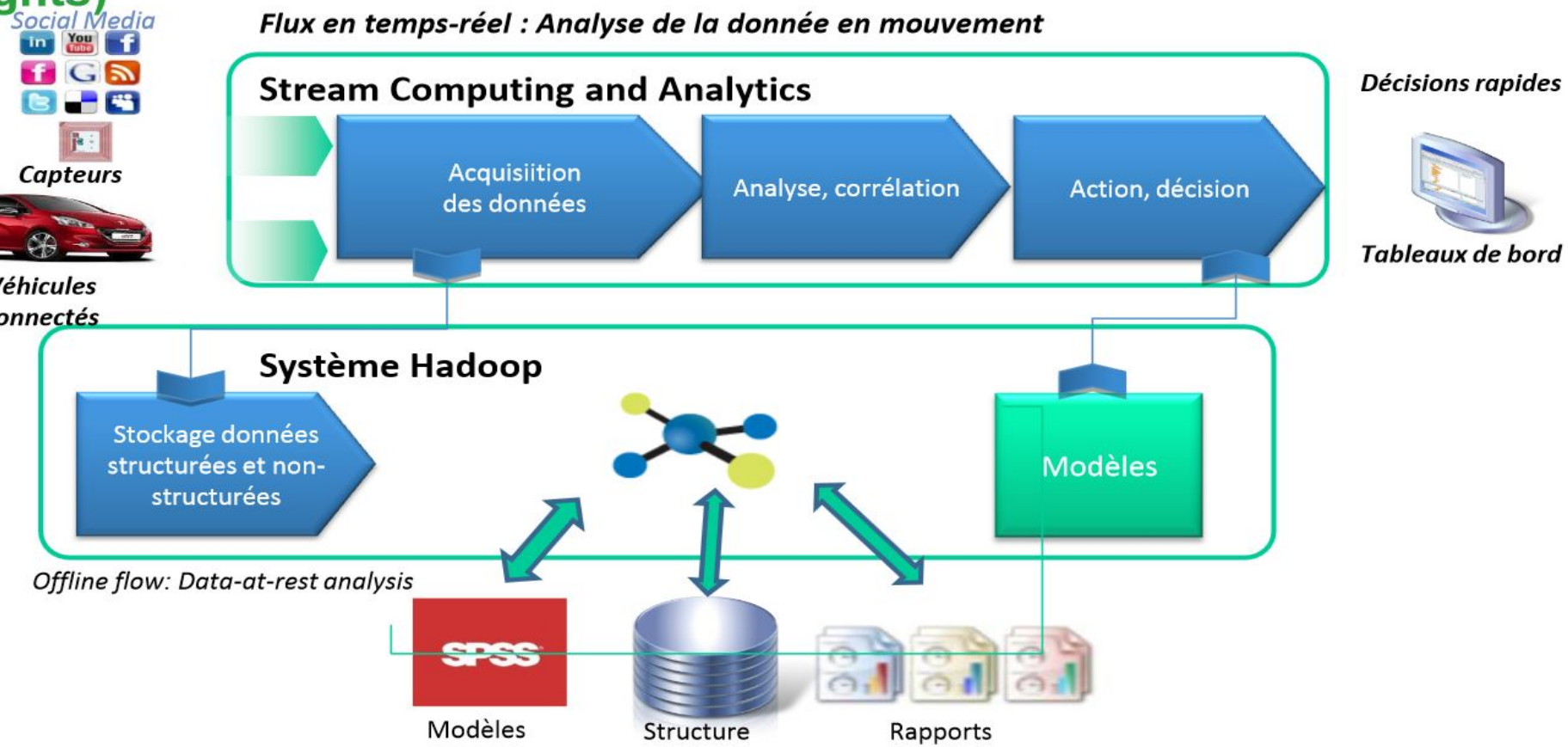
Market Manager Global Knowledge France

Agenda

- Présentation
- Big Data et Data Science
- Le métier du Data Scientist
- Les compétences du data scientist
- Cursus Global Knowledge
- Les profils recherchés par les entreprises
- Utilisation de SEM pour analyser le fichier

Plateforme IBM: Vehicule Connecté

Analyse de la donnée en mouvement (Streams) et stockage de la donnée (BigInsights)



Big Data et technologie

- Big Data concerne les ensembles de données très volumineux et les outils nécessaires pour les traiter et les analyser rapidement.
- Le Big Data a besoin d'une technologie pour:
  - Entrepôts de données de très grande capacité
  - Exécuter des requêtes analytiques sur d'énormes volumes de données structurées.
  - Moteur de traitements parallèles
  - La capacité d'analyser rapidement et agir sur les données tandis qu'elles sont toujours en mouvement .

Usage du Big Data par le Data Scientist

Résoudre les problèmes

Tout savoir sur son client

Operation avec Zero-latency

Innover: nouveaux produits

Détecter la fraude

Exploiter les objets et batiments connectés

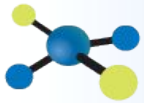
Capabilités

Visualisation et Découverte

Hadoop

Data Warehousing

Stream Computing



Integration et Gouvernance

Text Analytics

Analyser toutes les données

Transactional / Application Data

Machine Data

Social Media Data

Données

Business Analytics et Data Science

Les outils de Business Analytics permettent:

- Reporting et tableau de bord qui permettent de connaître l'activité actuelle et passée de l'entreprise;
- Visualisation qui diffèrent du reporting par l'animation et la possibilité de représenter l'effet d'un changement avec des graphes et des cartes de chaleur (heat Maps), des diagrammes de connection..
- Analyse prédictives et de sentiments qui permettent de comprendre l'activité actuelle de l'entreprise et de l' anticiper.

Analyse Prédictive

- L'analyse prédictive, connue aussi sous l'appelation Data Mining, concerne la modélisation statistique et/ou automatique des données.
- L'analyse prédictive résoud trois grandes catégories de problèmes:
  - Classification supervisée
  - Classification non supervisée
  - Analyse d'association et de panier d'achat
- Il arrive que pour résoudre un problème on combine ces trois classes de modèles
- Les éditeurs : IBM SPSS, SAS Institute, Coheris SPAD, R..

Analyse du texte ou Text mining

L'analyse des données textuelle s'appelle aussi l'analyse des données non structurées comme:

- Analyse des documents récoltées par l'entreprise. Exemple les mains courantes saisies dans les postes de police
- Analyse des E-mails
- Analyse des fichiers Log
- Tweets
- Messages postés sur Facebook

Les grand constructeurs de logiciels d'analyse prédictive se sont aussi dotés de logiciels d'analyse du texte.

La Méthodologie pour l'analyse prédictive

Compréhension de la problématique	Compréhension des données	Préparation des données	Modélisation	Evaluation	Déploiement
Déterminer les objectifs *Critères de réussite* **Situation** **Ressources** *Risques* *Couts et bénéfices attendus* **Déterminer les objectifs du DM** *Critères de réussite* **Planification des tâches**	**Collecter les données** **Analyse descriptive** **Vérifier la qualité des données**	**Sélectionner des données** *Inclusion/Exclusion* **Nettoyer les données** **Construire de nouveaux agrégats** **Reformater les données** **Construction des échantillons**	**Sélectionner les techniques de Modélisation** *Supervisé* *Non-supervisé* **Construire les modèles** *Choix des paramètres* *Description des modèles*	**Evaluation des résultats** *Critères de réussite* *Choix des modèles* **Revoir le process** *Retourner à l'étape de préparation des données* **Déterminer les prochaines étapes** *Liste des actions possibles* *Décision*	**Plan de déploiement** **Maintenance du déploiement** **Production du rapport final** *Rapport final* *Présentation finale*

- **CRISP-DM** (**CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining, http://www.crisp-dm.org/).

Les compétences recquises

- Analyse du problème et sa conception.
- Acquisition des données structurées et non structurées
- Préparation des données
- Modélisation statistique et machine learning
- Compétence de logiciels d'analyse prédictive tels que SAS, IBM SPSS
- Compétence en programmation avec R, Python,..
- Des connaissances ou expérience de Hadoop/mapReduce, Java

Qu’est-ce qu’une formation diplômante RNCP?

- **Les certifications répertoriées au RNCP identifient les formations professionnalisantes qui permettent d'obtenir un titre certifié reconnu par l'état et le Ministère du Travail sur l'ensemble du territoire national**
- **Un titre RNCP reconnaît un niveau de compétences**
- **Un diplôme valide un niveau d'études et d'éducation**

Equivalences	
Niveau I	niveau de formation supérieur à celui de la maîtrise (Bac +5)
Niveau II	niveau de formation équivalent à la licence ou la maîtrise (Bac +3/4)
Niveau III	niveau de formation équivalent à un DUT, à un BTS ou à une fin de premier cycle de l'enseignement supérieur (Bac +2)
Niveau IV	niveau de formation équivalent à un BP (brevet professionnel), à un BT (brevet de technicien), au Bac Professionnel ou technologique
Niveau V	niveau de formation équivalent à un BEP, à un CAP ou à un CFPA.

Cursus « Data Scientist » de Global Knowledge

Nous avons fait 2 cursus:

- Le premier très opérationnel correspond à 24 jours.
- Le deuxième plus long donne les bases pratiques et théoriques et bénéficie du label RNCP niveau I.

➢ Use Cases en travaux pratiques: Marketing, Risque bancaire, Maintenance Prédictive, Analyse d'opinion

- Permet les certifications IBM SPSS Modeler et Statistics:
  - ➢ IBM Certified Associate - SPSS Modeler Data Analysis
  - ➢ IBM Certified Associate - SPSS Modeler Data Mining
  - ➢ IBM Certified Specialist - SPSS Modeler Professional
  - ➢ IBM SPSS Statistics level 1

Cursus « Data Scientist » de Global Knowledge

- Le contenu des 2 cursus comprend :
  - Introduction aux technologies du Big Data
  - Présentation et visualisation des données
  - Analyse des données et Modélisation avec IBM SPSS Statistics et Modeler
    - Modèles Statistiques et Machine Learning
  - Programmation dans les outils IBM SPSS et avec R
  - Use Cases en travaux pratiques: Marketing, Risque bancaire, Maintenance Prédictive, Analyse d'opinion

- Le plus du cursus RNCP les rappels théoriques de base de la théorie des probabilités et statistiques, une plus grande diversité de modèles par problématique ainsi que des présentations théoriques plus approfondies. Et les « Uses Cases » sont plus développés.

Offre d'emploi: Data Scientist (Financial Services)

Vous avez une formation supérieure Bac 5 scientifique (type ENSAE, ENSAI, X, …) . Vous avez une expérience double en statistiques/modélisation et technologie et de solides connaissances des produits d'assurance ou financiers et des langages informatiques.

Description du poste:

- Savoir analyser l'ensemble des données internes et externes à disposition de l'entreprise.
- Permettre une exploitation optimale de la « connaissance client » dont dispose l'entreprise.
- Exploiter et croiser l'ensemble des données à disposition de la compagnie (site internet, carte bancaire, recouvrement, …).

Offre d’emploi: Data Scientist

- Vous travaillez sur des algorithmes d'apprentissage machine innovantes pour améliorer la plate-forme de mon client.

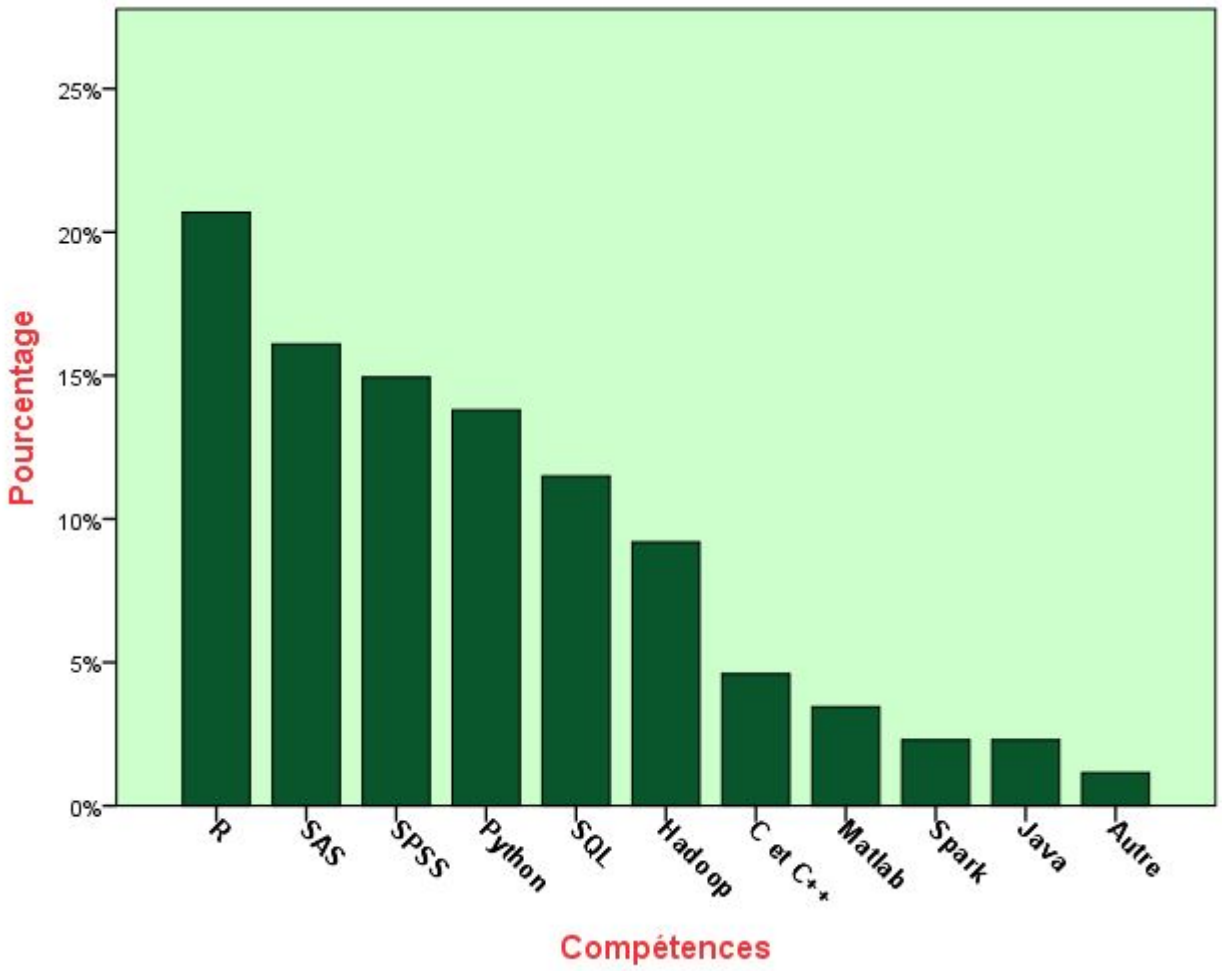
  Les principaux domaines de recherche sont la détection automatisée des anomalies et de prévision dans un environnement des offres en temps réel.

- Vos principales fonctions comprennent :

  l'analyse des données, la conception, la construction et la mise en œuvre de nouveaux algorithmes.

Vous devez avoir au moins 5 ans d'expérience avec l'analyse statistique. Les mathématiques appliquées mais aussi SQL MongoDB n’ont aucun secret pour vous

Echantillon: Site de l'Apec
Sur les 60 dernières offres du 16/11/2015



Analyse « Incidence »: DataSetEns

- Préparation des données:
  - Cancer = C18 (Cancer du Colon)
  - Par Age Sex et Pays: (x-min)/(max-min)
  - Structuration du fichier

- Comparaison de courbes de croissance par SEM pour France et US par Classe d'age et Sex.

Global Knowledge