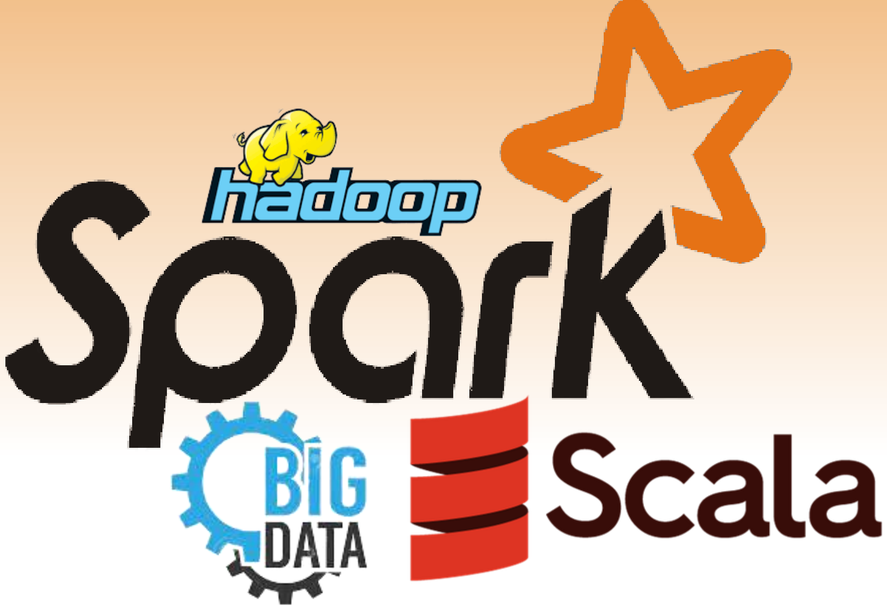


Initiation au Big Data
et
traitement de données massives avec Apache Spark

A propos de Moi

2017-2019:
- Msc en Big Data, AIMS Mbour
- Msc en Algèbre appliquée à la Cryptographie, Université de Thiés

2019-2020: Big Data Engineer @ MNS Consulting
- Installation d'un environnement Big Data et Mis en place d'une plateforme d'analyse et de prise de décision, se basant sur les données mobile.
- Mis en place de plateforme de visualisation des données provenant des capteurs IOT's.

2020-Present: Big Data Engineer @ Atos Sénégal
- Aider les métiers sur le développement des KPI's.
- Mise en place de pipeline d'ingestion de données pour alimenter le Datalake
- Mettre à la disposition des data scientists, les données agrégées.

Ibrahima FALL
Lead Big Data Engineer @ Atos Senegal,
Email: ibrahima.fall@atos.net
Alt-email: iboudofall@gmail.com

Sommaire

1. Le Big Data et ses caractéristiques
2. Hadoop et son écosystème
3. Initiation à Spark
    - Architecture
    - Spark API's
    - Spark operations
    - Exécution Plan
    - Demo

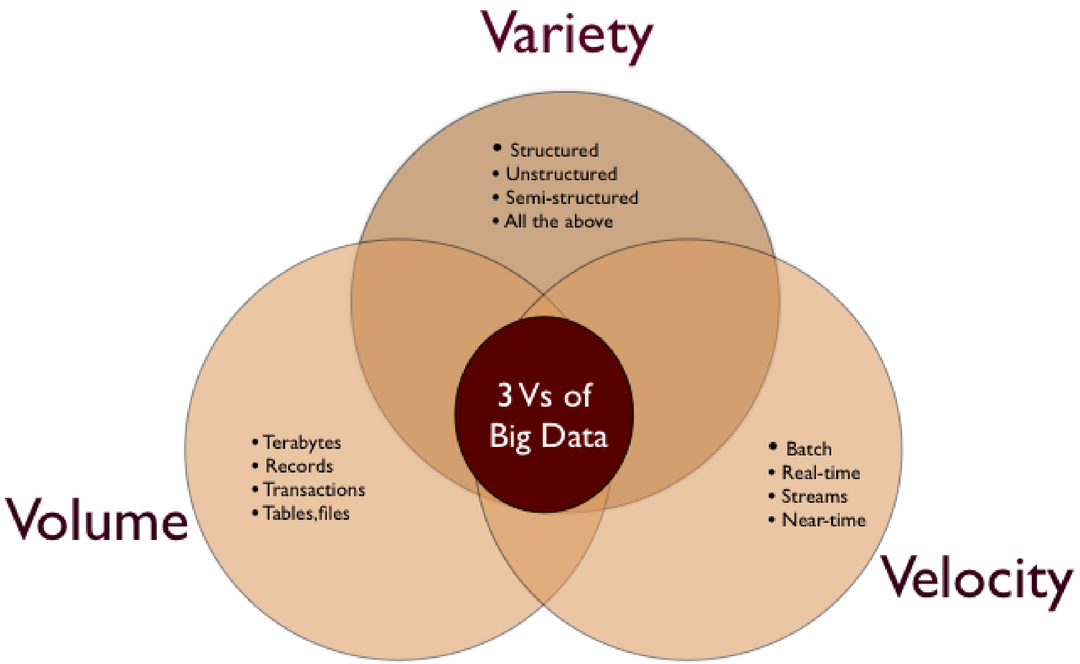
Definitions

What is Data ?

Definitions

What is Big Data ?

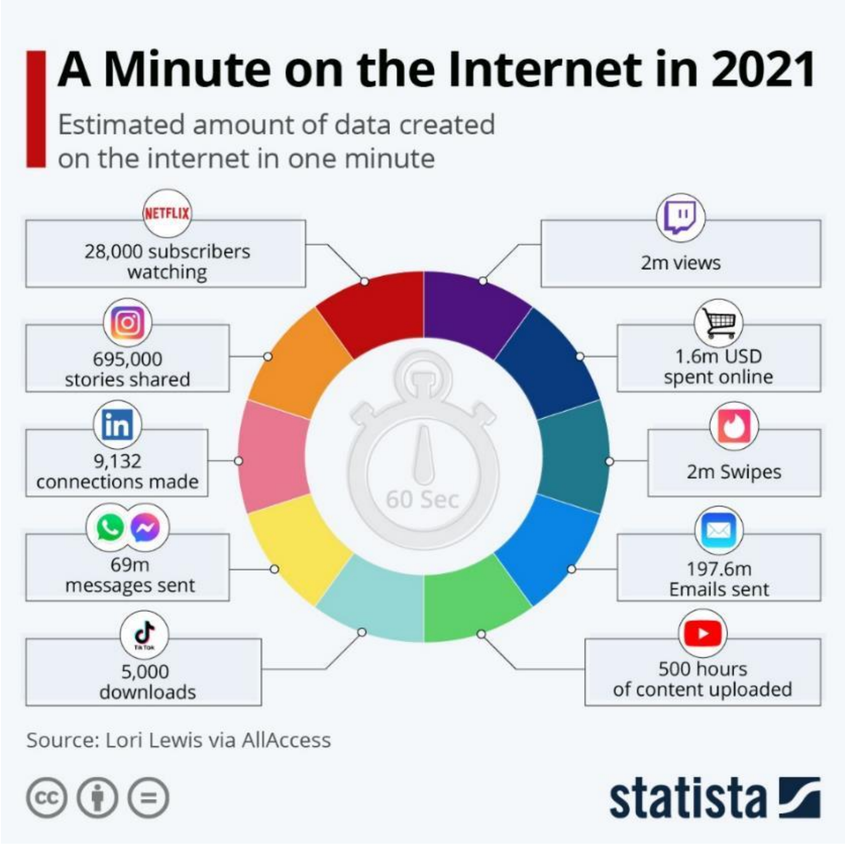
Data with specific characteristics



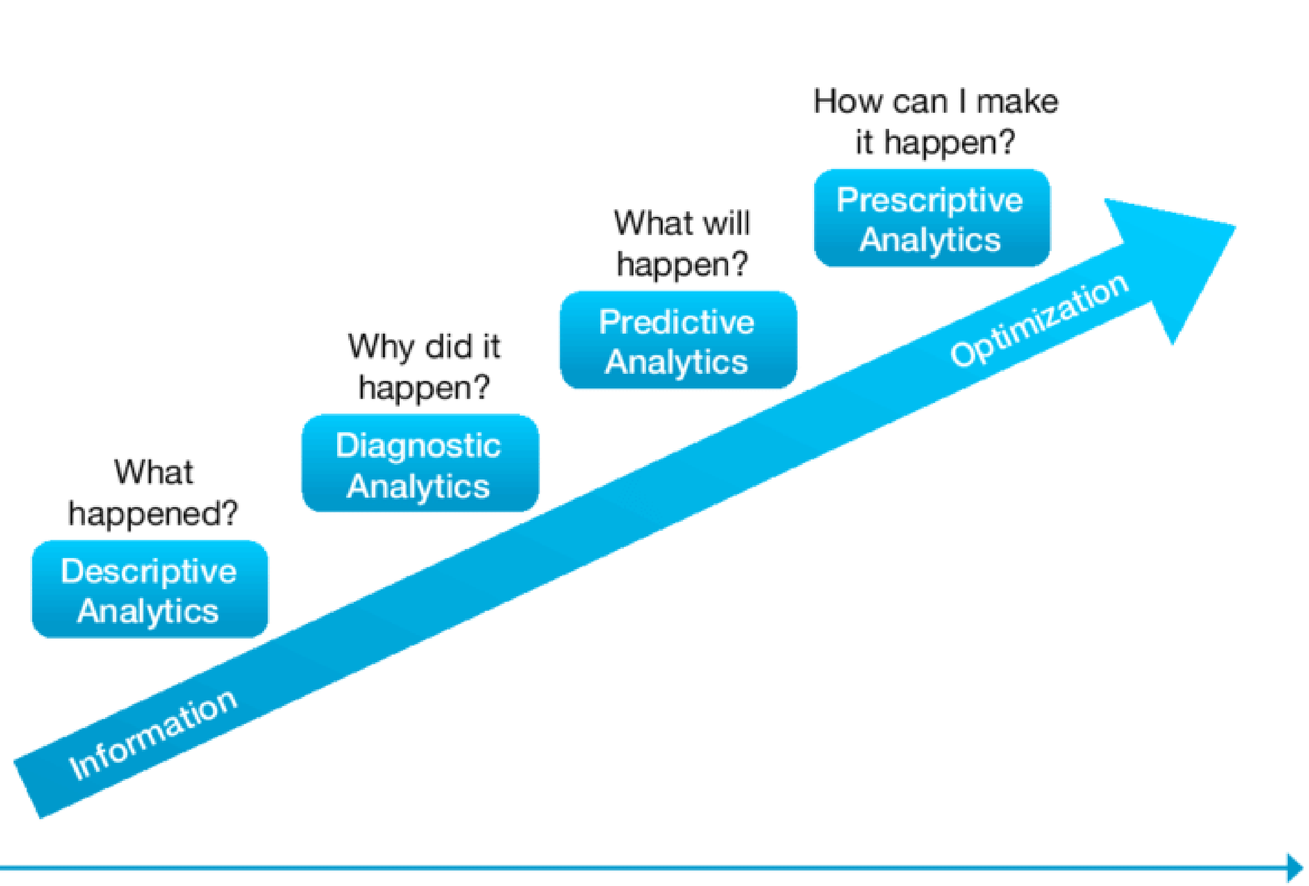
Technology, Strategy, Technics

Fast Growing

- Facilité de collecter des données grâce à une automatisation des systèmes.
- Des capteurs omniprésents pour générer des données
  - Téléphone portable
  - Logs des logiciels
  - Caméras
  - Capteurs IOT
  - …
- Le stockage devient accessible
- Prise de conscience que les données étaient très précieuses pour être supprimées



Type of analysis



Data Lifecycle

Big Data Jobs

Data Scientist

La science des données est une méthode systématique dédiée à la connaissance, découverte via l'analyse des données

- En entreprise, optimiser les processus organisationnels pour Efficacité
- En sciences, analyser des données expérimentales/observationnelles pour
- dériver des résultats

Skills:

- Computer Science
- Statistiques + Mathématiques
- Programmation: Python, R, SAS, JAVA, SQL
- Machine Learning
- Visualisation
- Connaissances du domaine

Big Data Jobs

Data Engineer

L'ingénierie des données est le domaine qui développe et fournit

- Des systèmes de gestion et d'analyse de données volumineuses
- Construire des plates-formes de données modulaires et évolutives pour les données scientifiques
- Déployer des solutions Big Data

Skills:

- Computer Science
- Bases de données
- Traitement de données en temps réel / données massives
- Programmation: Python, Scala, SQL, …
- Comprendre les facteurs de performance et les limites des systèmes

Big Data Jobs

Data Analyst

Analyser les données pour les transformer en informations exploitables

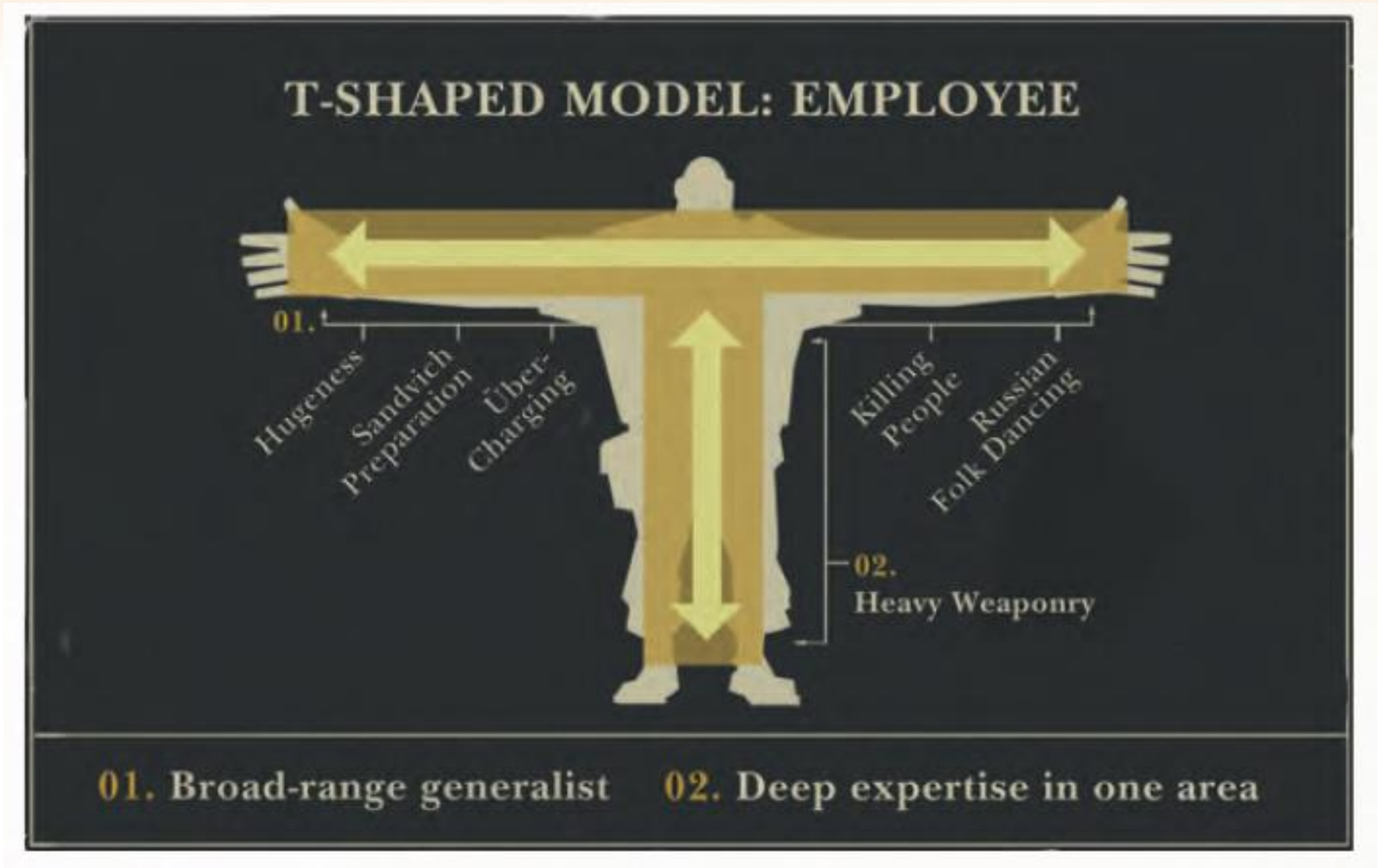
- Définir la stratégie Data-Driven de l'entreprise
- Créer et maintenir les bases de données de l'entreprise
- Elaborer les critères de segmentation

Skills:

- Computer Science
- Statistiques
- Utilisation Excel
- Visualisation de données
- Programmation: Python, Scala, SAS, SQL, …

Data Engineer Vs Others

Be T-shaped


T-SHAPED MODEL: EMPLOYEE
01.
Hugeness
Sandvich Preparation
Über-Charging
Killing People
Russian Folk Dancing
02.
Heavy Weaponry
01. Broad-range generalist
02. Deep expertise in one area

Use Cases

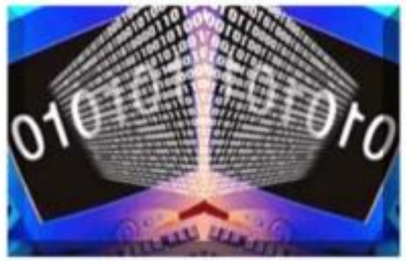
BIG DATA

Smarter Healthcare

Multi-channel

Finance

Log Analysis



Homeland Security

Traffic Control

Telecom

Search Quality

Manufacturing

Trading Analytics

Fraud and Risk

Retail: Churn, NBO

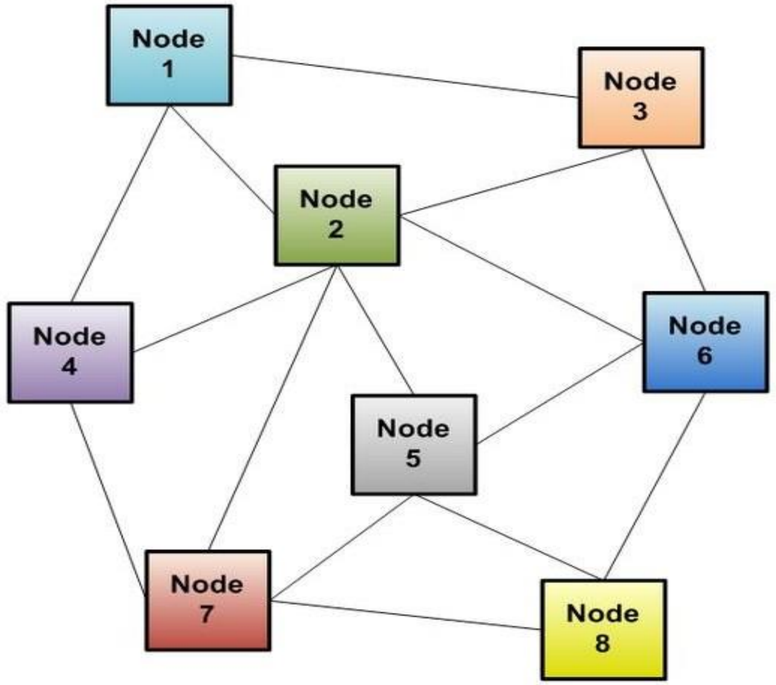
Big data Challenges

Distributed System

Un système distribué est un environnement informatique dans lequel divers composants sont répartis sur plusieurs ordinateurs (ou autres appareils informatiques) sur un réseau. Ces appareils divisent le travail, coordonnent leurs efforts pour accomplir le travail plus efficacement.

Caractéristiques:

- Scalabilité
- Concurrence
- Disponibilité / fault tolerance
- Transparence
- Hétérogénéité
- Réplication
- Géo-distribution



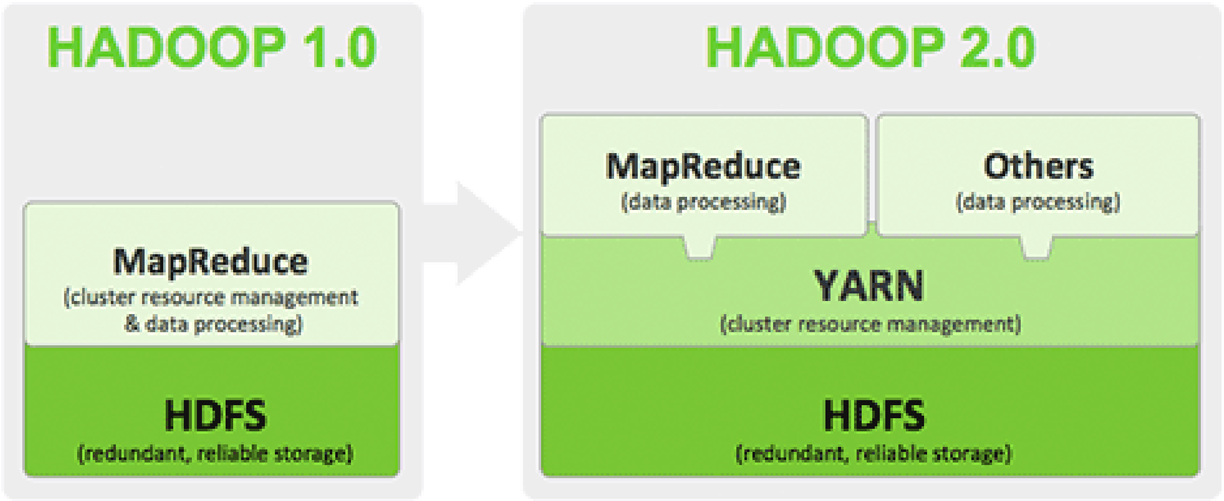
What is Hadoop

Apache Hadoop

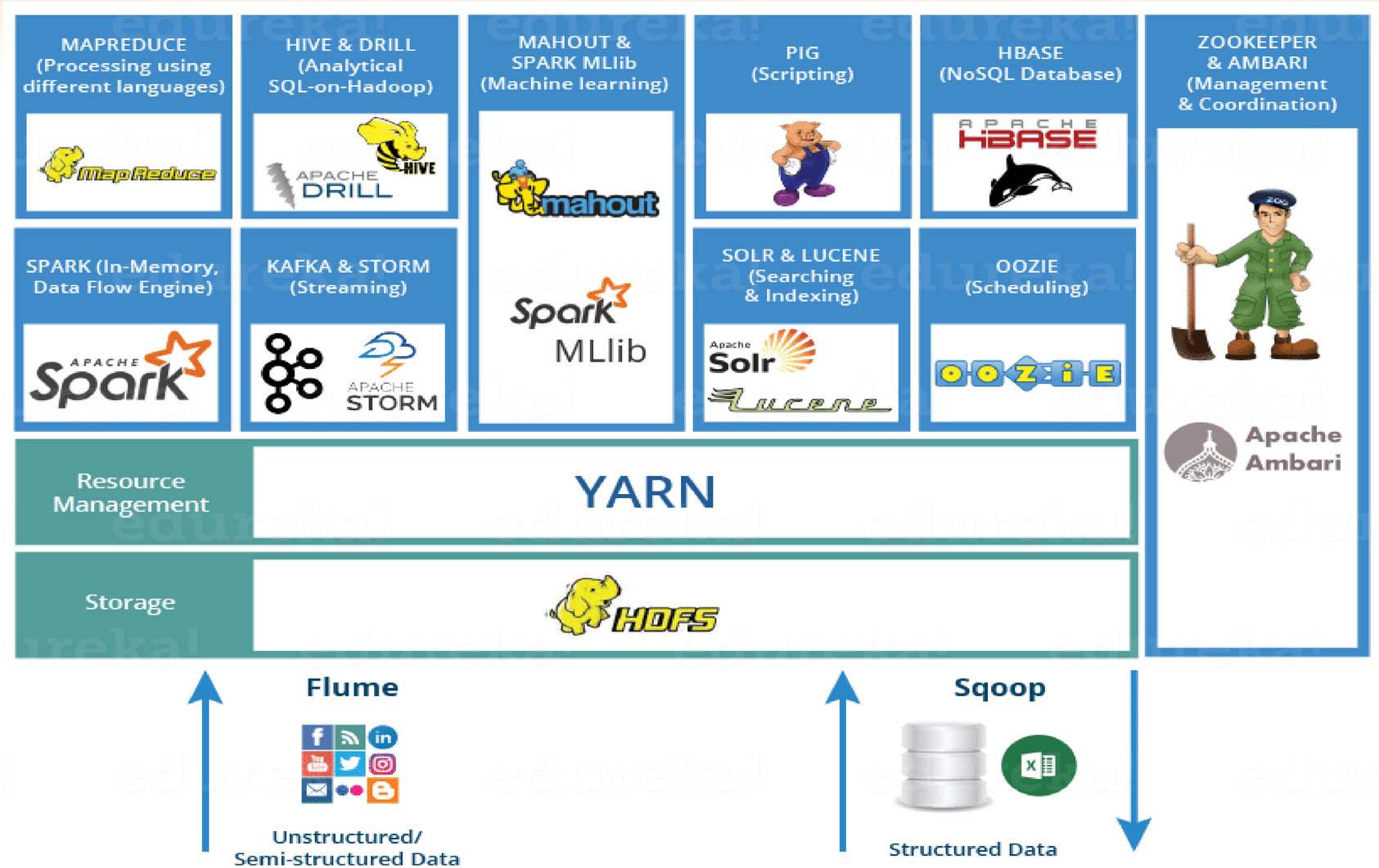
The Apache™ Hadoop® project develops open-source software for reliable, scalable, distributed computing.

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.

- Hadoop est une plateforme Big Data open-source, utilisée pour stocker et traiter d'immenses volumes de données d'une manière distribuée.
- Composants principaux:
  - HDFS (Hadoop Distributed File System)
  - YARN (Yet Another Ressource Negociator)
  - MapReduce



Hadoop Ecosystem



What is Spark?

Apache Spark est un moteur de traitement de données ,

- in-memory
- open-source
- ultra-rapide
- tolérant aux pannes
- supporte plusieurs languages
- Moteur unifié
- lazy computations

pour le big data et le machine learning, avec des modules pour des traitements de données en temps reel, sql et calcul de graph.

APACHE Spark™ *Lightning-fast unified analytics engine*

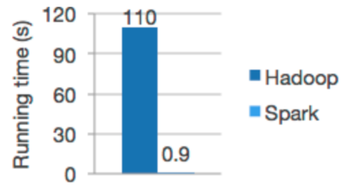
Download | Libraries | Documentation | Examples | Community | Developers

Apache Spark™ is a unified analytics engine for large-scale data processing.

Speed

Run workloads 100x faster.

Apache Spark achieves high performance for both batch and streaming data, using a state-of-the-art DAG scheduler, a query optimizer, and a physical execution engine.



Logistic regression in Hadoop and Spark

Ease of Use

Write applications quickly in Java, Scala, Python, R, and SQL.

Spark offers over 80 high-level operators that make it easy to build parallel apps. And you can use it *interactively* from the Scala, Python, R, and SQL shells.

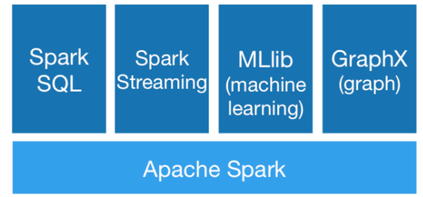
```
df = spark.read.json("logs.json")
df.where("age > 21")
  .select("name.first").show()
```

Spark's Python DataFrame API
Read JSON files with automatic schema inference

Generality

Combine SQL, streaming, and complex analytics.

Spark powers a stack of libraries including SQL and DataFrames, MLlib for machine learning, GraphX, and Spark Streaming. You can combine these libraries seamlessly in the same application.



Spark vs Hadoop ?

- En réalité, Spark et comparable à MapReduce et non à Hadoop.
- Spark peut se branché à Hadoop et utiliser HDFS pour bénéficier du système de stockage de cette dernière.
- Spark est plus rapide que MapReduce (10 à 100 fois)
- Spark est écrit en Scala, MapReduce est écrit en Java
- En traitement, Spark supporte du **batch**/ **real-time**/ **graph** et MapReduce ne supporte que du **batch**
- MapReduce ne supporte pas du Caching, contrairement à Spark

Why Spark ?

Rapidité

Facilte le développement avec les api Java, Scala, Python, R, SQL

Flexibilité

Syntaxes trés simples

Moteur unifié

Apache Spark — Lightning-fast unified analytics engine

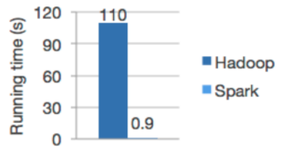
Download | Libraries | Documentation | Examples | Community | Developers

Apache Spark™ is a unified analytics engine for large-scale data processing.

Speed

Run workloads 100x faster.

Apache Spark achieves high performance for both batch and streaming data, using a state-of-the-art DAG scheduler, a query optimizer, and a physical execution engine.



Logistic regression in Hadoop and Spark

Ease of Use

Write applications quickly in Java, Scala, Python, R, and SQL.

Spark offers over 80 high-level operators that make it easy to build parallel apps. And you can use it *interactively* from the Scala, Python, R, and SQL shells.

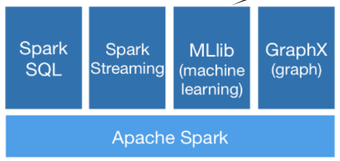
```
df = spark.read.json("logs.json")
df.where("age > 21")
  .select("name.first").show()
```

Spark's Python DataFrame API
Read JSON files with automatic schema inference

Generality

Combine SQL, streaming, and complex analytics.

Spark powers a stack of libraries including SQL and DataFrames, MLlib for machine learning, GraphX, and Spark Streaming. You can combine these libraries seamlessly in the same application.



Runs Everywhere

Spark runs on Hadoop, Apache Mesos, Kubernetes, standalone, or in the cloud. It can access diverse data sources.

You can run Spark using its standalone cluster mode, on EC2, on Hadoop YARN, on Mesos, or on Kubernetes. Access data in HDFS, Alluxio, Apache Cassandra, Apache HBase, Apache Hive, and hundreds of other data sources.

Apache Spark, hadoop, cassandra, MESOS, APACHE HBASE, kubernetes

Use of Spark

Suivant le projet, Spark peut être utilisé

- En mode single node

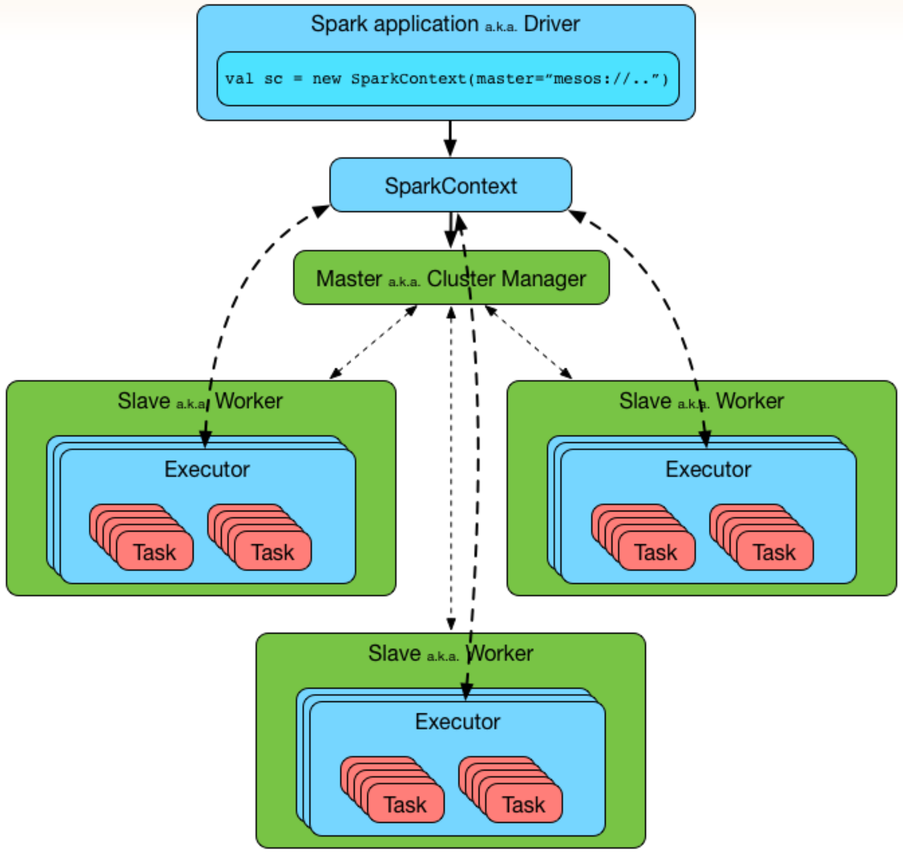
- En mode cluster, sur plusieurs machines

  - On Premise

  - On the cloud: Amazon Web Services(AWS), Microsoft Azure

Spark Architecture

- Driver : c'est le processus contenant le programme spark, il crée le DAG (Directed Acyclic Graph), planifie et coordonne l'exécution du programme
- SparkContext (SparkSession) : il représente la porte d'entrée du Spark Cluster avec des configurations prédéfinies (Master, nom de l'application, mémoire, nombre d'exécuteurs …)
- Cluster Manager : négocie de la ressource et la rend disponible pour le driver, suit le statut des workers.
- Executors : lieu d'exécutionn des tasks du DAG



Source: https://sunbiaobiao.gitbooks.io/sa/content/spark-architecture.html

Master et le Cluster Manger peuvent être fusionnés comme en mode StandAlone ou séparés comme en mode YARN

Spark API's

Spark APIs: RDD, Dataset et DataFrame

RDD

- RDD = Resilient Distributed Dataset (jeux de données distribué et resilient)
- RDDs sont les structures fondamentales de Spark
- RDDs sont tolerant aux pannes, immuables, partitionnés, distribués, supportent le lazy evaluation
- Compile-time Type Safe

DataFrames

- Même structure que pandas dataframe et dataframes dans R
- Représente une abstraction des RDDs
- Supporte Catalyst Optimisation
- Erreurs syntaxiques décelées à la compilation
- Erreurs analytiques décelées au runtime

Dataset

- Un union de RDD et DataFrame

Spark Operations

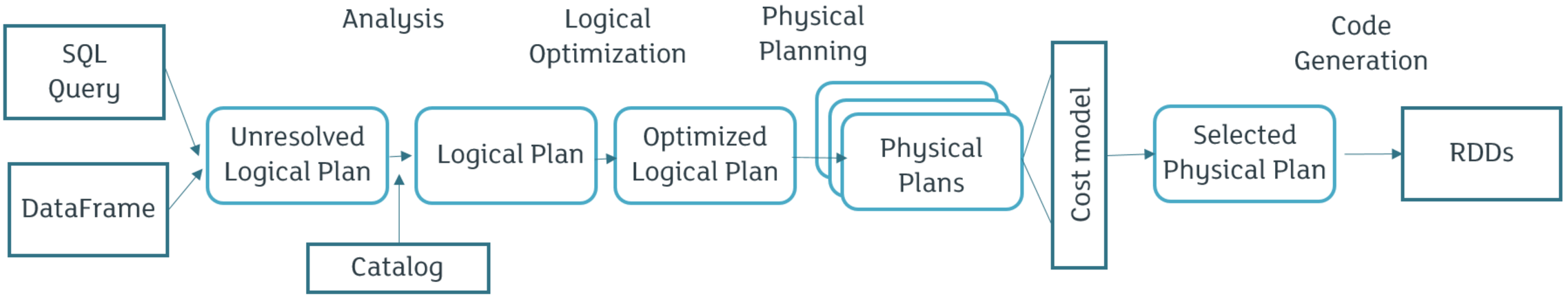
Spark dispose de deux types d'operations:

- Transformations: créer une nouvelle dataset à partir d'une autre existente
- Actions: retoune une valeur finale ou écrit le résultat dans un espace de stockage externe.

Les transformations sont dites "lazy": elles ne s'executent pas immédiatement: en effet, elles gardent juste l'ensemble des operations qui ont permit de construire la dataset résultante (lineage). Les transformations ne sont executées que lorsqu'une action est réquise.

L'ensemble des operations faites sur un dataset forme un DAG (Directed Acyclic Graph) dont les sommets sont des RDD et les arêtes représentent les operations.

Plan d'exécution



Source: https://medium.com/codex/apache-spark-optimization-techniques-and-tuning-52cc446c07f5

Enough Talk Let's Code