Methodology for Big Data Project*

1. Preparation
2. Strategy
3. Governance
4. Architecture
5. Best pratices

*Main source: Le grand manuel des big data. Informatica 2017.

1. Preparation

- Without preparation, too many big data projects fail
  - Objectives too vague or too ambitious
  - Expectations too high, even unrealistic
  - Exceeding budget and deadlines
  - Inability to evolve over the long term
- A crucial step
  - Set clear and precise objectives, in order to demonstrate the value of the project to business users
  - Define indicators to measure the success of the project
  - Look for the right tools to increase development productivity

Choosing the Project

- A tactical project, for a need and a service
  - Can then be adapted for other services
- Characteristics of the project
  - Demonstrate its value to the business service
  - Support at the highest level and sharing of the vision
  - Value easily transferable within the company to other departments
    - E.g. moving from marketing to logistics
  - Acquisition of transferable skills and lessons

2. Strategy

- Definition of business objectives
  - Specific, e.g. reduce monthly customer loss by 20%
  - Duration to achieve it, e.g. 3 to 6 months
- Definition of IT objectives
  - In support of business objectives, e.g. setting up a customer data integration process
  - Duration and measure of success, e. g. 90% prediction rate of customer loss
- Definition of data requirements
  - Identify the necessary data, internally (e.g. dormant data in silos) and externally
  - Characterize the data
    - Volume, variety, velocity, veracity
  - Compliance with standards: security, privacy,...

Team

- Complementarity between understanding business objectives and technical aspects
- Technical skills
  - Properly identify the need for new skills (e.g. Hadoop) and the integration of new recruits
  - Use the skills for which employees have been hired to avoid demotivation or departure
  - Anticipate the evolution of skills
  - Be careful not to want to code everything in Java in Hadoop
  - Beware of the NIH (not invented here) of the GAFAM

Tools

- Understand and master the tools
  - Data analysis
  - Machine learning
  - Visualization
  - Storage in files or databases
  - Cluster management
  - Integration, ETL, data lake
  - Etc.
- Properly assess their maturity (POC)
  - Many bugs in new products

3. Data Governance

- When?
  - The big data project is company-wide (involves several divisions or departments)
  - To move from POC to industrialization
- Why?
  - Develop the business (data quality)
  - Saving and rationalizing
  - Increase agility and productivity
  - Comply with the law (e.g. GDPR)
- How?
  - Data Governance Committee, responsible for overseeing the company's data policy
  - In connection with data stewards

Processus

- Implementation of efficient, reusable and scalable processes for the following steps
  - Data access (streaming, extraction,...)
  - Integration of various data
  - Cleaning (duplicate removal, error correction,...)
  - Data control (consolidation, enrichment,...)
  - Securing (e. g. masking sensitive data)
  - Data analysis
  - Business requirements analysis
  - Use of information

4. Architecture

- Start small with a sandbox
  - Well controlled environment, e.g. Spark/Hadoop on a server
  - Plan for scale-up, e.g. distribution on n servers
  - Hide test data from production
  - Correct coding errors
- Switch to the target architecture
  - E.g. data lake with data ingestion process and data delivery to applications

5. Best Practices

- Start with clear, measurable objectives
- Have the support of the business
- Ensure that data is accessible
- Build a team with business and big data skills
  - Hire data scientists versus train BI experts
- Establish governance
- Start small
- Seeing far away
  - Plan for scaling up in volume and load

For More Information

Some MOOCs

- Big Data Specialization at Coursera
  - https://fr.coursera.org/specializations/big-data
  - Introduction to Big Data
  - Hadoop Platform and Application Framework
  - Introduction to Big Data Analytics
  - Machine Learning With Big Data
  - Graph Analytics for Big Data
- Big Data fundamentals
  - https://www.fun-mooc.fr/courses/MinesTelecom/04006S04/session04/about
  - Data management, statistics and machine learning
- Understand Big Data through movies
  - https://openclassrooms.com/courses/comprendre-le-big-data-a-travers-les-films-de-cinema
  - Basic and fun
- Data lake
  - https://educast.emc.com/learn/data-lakes-for-big-data

Some Top Web Sites

- apache.org
- hadoop.apache.org
- spark.apache.org
- big-data.developpez.com
  - French forum for developers
- nosql.developpez.com/cours
  - NoSQL tutorials
- bigdatauniversity.com
  - A big data "university", open and free
- bigdata-madesimple.com
  - Portal to many big data ressources
- dataconomy.com/big-data-blogs
  - The top blogs