Présentation & Subject Analysis

Identification :

Project Title	POC datawarehouse big data
Project Number	203
Team Referent	GREGORI Anthony
Other members in team	AISSAOUI Ramez, HILALI Elies, STOJANOVIC Maxime
Team Mentor	
Project Partner	Acensi

1 Context

We are with acensi a company created 13 years ago, specialized in industry council in computer software and systems. We will work with the technical director and some people on a project in the continuation of a previous one which is a Datawarehouse. The current project use Microsoft SSRS (SQL Server Reporting Services) on windows 2012 R2, MariaDB and Talend on Debian.
The actual datawarehouse get information of:

- Microsoft AD (Azure Active Directory) is the Microsoft management service directories and identity base on shared cloud
- CRM – Customer Relationship Management: is the way to optimize the interact of enterprise with clients.
- Accounting software
- Ticketing tool is a way to follow some glitch
- Microsoft Office 365
- The extranet and intranet company
- Telephony (Orange & Bouygues)

The goal of the actual datawarehouse is to make functional relationships but also error reports, mainly on licensing issues and inconsistent data.

So know the company want a BigData solution, which complete the actual datawarehouse with some oders data like the API, applicative logs, CSV and other data. Moreover, if we have a better solution than actual, our solution could replace the old one.
So for the moment, we must learn what is the bigdata, which are the software useful, what are the current solution, what solution can we suggest, how it works, how bigdata emerged.

2 First State of the Art

Some solution being by giant
//description of some solution, what we have on internet

On big Data, we have so much software, language. So we have to learn about them.
We have NoSQL, it allows system to be clusterisable, they often are without patterns, without transactions, they are non-relational (not join) and much are in open source. Also we can speak about the MapReduce algorithm and framework Hadoop, that allows you to do some request in parallelize but you must to have data structured in key-values, MapReduce on big Data must be implemented on a cluster of servers. Hadoop is a framework in open source written in java.
They are several components of Apache Hadoop, Hadoop Distributed File System, MapReduce and YARN, Hbase (a NoSQL base), ZooKeeper, pig, hive (Pig and Hive are for reading of logs), Oozie planning of Hadoop job, Flume can aggregate in real time different streams of logs and Sqoop efficiently transfer large volumes of data between Hadoop and an RDBMS (Relational Database Management System). Major distributions Hadoop are: Cloudera, Hortonworks, MapR, Amazon Elastic MapReduce:

- Cloudera Solution, largely faithful except for the administration tools.

Cloudera is intended as the commercial Hadoop company.
Founded by Hadoop experts from Facebook, Google, Oracle and Yahoo.
If their platform is largely based on Hadoop Apache, it is complemented with homemade components primarily for cluster management.

Cloudera's business model is the sale of licenses but also the support and training.
Cloudera offers a fully open source version of their platform (Apache 2.0 License). To date, it is the distribution en Hadoop has the most referenced deployments.

- Solution Hortonworks, true to the Apache distribution and therefore 100% open source.

Hortonworks was formed in June 2011 by members of the team in charge of Yahoo Hadoop project.
Their goal is to facilitate the adoption of the Hadoop platform Apache, so all components are open source under the Apache license.

The economic model of Hortonworks is not to sell licenses but only support and training.
This distribution is the most consistent with the Hadoop platform and Apache is a big contributor Hortonworks Hadoop.

- Solution MapR Hadoop core but repackaged and enriched with proprietary solutions.

MapR was founded in 2009 by former members of Google.

Although his approach is commercial, MapR contributes to projects like Apache Hadoop HBase, Pig, Hive, and especially ZooKeeper Drill.

MapR stands above all the version of Apache Hadoop with its shooting distance with the heart of the platform. They put forward their own distributed file systems and their own versions of MapReduce.

The three distributions have a different approach and positioning regarding the vision of a Hadoop platform (open source business model ...).

The choice will be one solution or the other depending on the requirements:

* Open Source Solution.
* Maturity of the solution.
* Partnerships and compatibility with satellite products.

The choice of a distribution is more difficult that the future of Hadoop that is far from any trace.

3 Objectives and sub-objectives definitions

We must to learn what is the big data, we have to study the project, which solutions exist, which can be useful and how they do that. We have one month to study the project with company, the contest and offer a solution. After that we have one month on tests/POC (Proof Of Concept) and one month to really write the project so at this moment we should redo a schedule but more accurate than current.

4 Valorisation

Like we said, we have to learn so much about big data, a promising sector. We should be the first solution for big data in this company, and we are going to learn how manage a concrete project.
At the end our profiles can be save by the technical director, and we could be taken for a job.