1. Tìm hiểu các khái niệm và chức năng tổng quát của Hadoop, MapReduce, HDFS, Sqoop, Pig, Hive, Spark.

**Hadoop** is a collection of software utilities that employ a network of computers to solve problems using massive amounts of data and computation.

Hadoop has two major functions:

* It provides a distributed file system to big data sets.  
  Big data sets generally have the size of hundreds of gigabytes. For such huge data set it, Hadoop provides a distributed file system (HDFS), allowing data to be stored in clusters of different commodity machines and then be accessed parallelly.
* Thus transforming the data set into useful information using the MapReduce programming model.

**MapReduce** is a processing technique and a program model for distributed computing.

The MapReduce algorithm contains two important tasks, namely Map and Reduce.

* Map takes a set of data and coverts it into another set of data, where individual elements are broken down into tuples (key/value pairs).
* Reduce task takes the output from a map as an input and combines those data tuples into a smaller set of tuples.

**Hadoop Distributed File System (HDFS)** is a distributed file system designed to run on commodity hardware. HDFS is highly fault-tolerant and is designed to be deployed on low-cost hardware. HDFS provides high throughput access to application data and is suitable for applications that have large data sets.

Goals of HDFS

* Handling hardware failure. The HDFS contains multiple server machines. In case any machine fails, the HDFS will recover the node quickly.
* Streaming data access. The HDFS applications usually run on the general-purpose file system. This application requires streaming access to their data sets.
* Coherence Model. The application that runs on HDFS require to follow the write-once-read-many approach. So, a file once created need not to be changed. However, it can be appended and truncate.

**Sqoop** is a tool used for data transfer between RDBMS (like MySQL, Oracle SQL etc.) and Hadoop (Hive, HDFS, and HBASE etc.)

It is used to import data from RDBMS to Hadoop and export data from Hadoop to RDBMS.

**Apache Pig** is a high-level platform for creating programs that run on Apache Hadoop. Pig can execute its Hadoop jobs in MapReduce, Apache Tez, or Apache Spark. Pig abstracts the programming from the Java MapReduce idiom into a notation which makes MapReduce programming high level, similar to that of SQL for relational database management systems.

The applications of Apace pig are,

- It is used to process huge data sources like web logs, streaming online data etc.

- It Support Ad Hoc queries across large data set.

- Used to perform data processing in search platforms.

Example Yahoo uses Pig for 40% of their jobs including news feeds and search engine.

- It is also used to process time sensitive data loads. Example twitter data

i.e Analyze the data to find patterns in customer behaviors and make recommendations immediately like trending tweets.

- Apache Pig is generally used by data scientists for performing tasks like ad-hoc processing and quick prototyping.

**Apache Hive** is a data warehousing infrastructure based on Apache Hadoop. Hive is designed to enable easy data summarization, ad-hoc querying and analysis of large volumes of data. It provides SQL which enables users to do ad-hoc querying, summarization and data analysis easily. At the same time, Hive's SQL gives users multiple places to integrate their own functionality to do custom analysis, such as User Defined Functions (UDFs).

**Apache Spark** is a lightning-fast cluster computing technology, designed for fast computation. It is based on Hadoop MapReduce and it extends the MapReduce model to efficiently use it for more types of computations, which includes interactive queries and stream processing. The main feature of Spark is its in-memory cluster computing that increases the processing speed of an application.

1. So sánh kiến trúc các hệ thống trong slide: Facebook, Amazon, Reference.
2. 1) Facebook
3. Facebook collects data from two sources: a federated
4. MySQL tier containing user data; and web servers for event log
5. data. The structured data from the Federated MySQL is copied,
6. compressed and stored in the Production Hive-Hadoop cluster.
7. The Scribe servers aggregate event logs and process the data in
8. Hadoop Distributed File System (HDFS). HDFS data is
9. compressed periodically, and moved to Production Hive-
10. Hadoop clusters for further processing.

1) Facebook

Facebook collects data from two sources: a federated MySQL tier containing user data; and web servers for event log data. The structured data from the Federated MySQL is copied, compressed and stored in the Production Hive-Hadoop cluster. The Scribe servers aggregate event logs and process the data in Hadoop Distributed File System (HDFS). HDFS data is compressed periodically and moved to Production Hive-Hadoop clusters for further processing.

Facebook differentiates low or high priority clusters for

data analysis. High priority jobs are executed in the Production

Hive-Hadoop cluster whereas lower priority jobs and ad hoc

analysis jobs are executed in the Ad hoc Hive-Hadoop cluster.

In addition, data is replicated from the Production cluster to the

Ad hoc cluster. The results of data analysis are stored in the

Hive-Hadoop cluster or MySQL tier for Facebook users.

HiPal, a graphical user interface with Hive CLI (a Hive

command-line interface) is used to accommodate ad hoc

analysis. Databee (a python framework) is used for execution

and scheduling of periodic batch jobs in the production cluster.

Microstrategy Business Intelligence (BI) tools are used for

cube dimensional analysis.

The mapping Facebook use case reference architecture is

presented in Fig. 1.

2) Reference Architecture

From the mapping with each big data use case, we

compiled a comparison table in order to produce the reference

architecture for Big Data system. The comparison table is

presented in Table 1 as well as the reference architecture at Fig.

2.

3) Amazon Web Services with Reference Architecture

We also mapped our reference architecture with Amazon

Web Services [37]. At this stage, we do not target for a specific

requirement, business needs or use case, but rather focus on

evaluating our produced architecture with a commercial

service, Amazon. Fig. 3 shows the corresponding mapping of

the reference architecture.

Facebook differentiates low or high priority clusters for

data analysis. High priority jobs are executed in the Production

Hive-Hadoop cluster whereas lower priority jobs and ad hoc

analysis jobs are executed in the Ad hoc Hive-Hadoop cluster.

In addition, data is replicated from the Production cluster to the

Ad hoc cluster. The results of data analysis are stored in the

Hive-Hadoop cluster or MySQL tier for Facebook users.

HiPal, a graphical user interface with Hive CLI (a Hive

command-line interface) is used to accommodate ad hoc

analysis. Databee (a python framework) is used for execution

and scheduling of periodic batch jobs in the production cluster.

Microstrategy Business Intelligence (BI) tools are used for

cube dimensional analysis.

The mapping Facebook use case reference architecture is

presented in Fig. 1.

2) Reference Architecture

From the mapping with each big data use case, we

compiled a comparison table in order to produce the reference

architecture for Big Data system. The comparison table is

presented in Table 1 as well as the reference architecture at Fig.

2.

3) Amazon Web Services with Reference Architecture

We also mapped our reference architecture with Amazon

Web Services [37]. At this stage, we do not target for a specific

requirement, business needs or use case, but rather focus on

evaluating our produced architecture with a commercial

service, Amazon. Fig. 3 shows the corresponding mapping of

the reference architecture.

Facebook differentiates low or high priority clusters for data analysis. High priority jobs are executed in the Production Hive-Hadoop cluster whereas lower priority jobs and ad hoc analysis jobs are executed in the Ad hoc Hive-Hadoop cluster. In addition, data is replicated from the Production cluster to the Ad hoc cluster. The results of data analysis are stored in the Hive-Hadoop cluster or MySQL tier for Facebook users. HiPal, a graphical user interface with Hive CLI (a Hive command-line interface) is used to accommodate ad hoc analysis. Databee (a python framework) is used for execution and scheduling of periodic batch jobs in the production cluster. Microstrategy Business Intelligence (BI) tools are used for cube dimensional analysis. The mapping Facebook use case reference architecture is presented in Fig. 1.

2) Reference Architecture

From the mapping with each big data use case, we compiled a comparison table in order to produce the reference architecture for Big Data system. The comparison table is presented in Table 1 as well as the reference architecture at Fig. 2.



