Project Vodafone USA (ATLAS)– Telecom Industry

In this slide we will be discussing about a data analysis and data mining tools used in

Vodafone – Telecom Industry project .

Phases of project :

There are three phases in the above given project.

1) Data migration.
2) Data processing.
3) Data analysis.

Data migration :

Data migration is the process of transferring **data** between computer storage types or file formats. It is a key consideration for any system implementation, upgrade, or consolidation. **Data migration** is performed to achieve an automated **migration**, freeing up human resources from tedious tasks.

Data migration is a phrase used to describe the process of moving data from one storage device to another. In this context, data migration is the same as Hierarchical Storage Management (HSM).

Data Processing:

Data processing is, generally, "the collection and manipulation of items of data to produce meaningful information."[1] In this sense it can be considered a subset of *information processing*, "the change (processing) of information in any manner detectable by an observer." [note 1]

The term Data Processing (DP) has also been used to refer to a department within an organization responsible for the operation of data processing applications.

Data analysis :

Data analysis is a process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making.

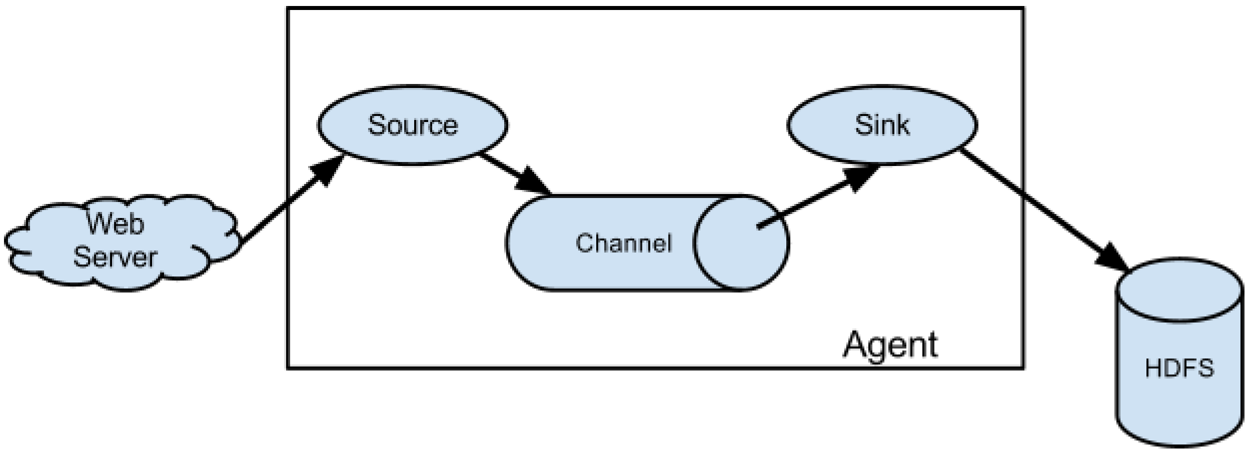
TOOLS USED IN THIS PROJECT

- SQOOP
- FLUME
- HIVE
- PIG

What is Apache Flume?

Apache Flume is a Hadoop ecosystem component used to collect, aggregate and moves a large amount of log data from different sources to a centralized data store.
It is an open source component which is designed to locate and store the data in a distributed environment and collects the data as per the specified input key(s).

A Flume event is defined as a unit of data flow having a byte payload and an optional set of string attributes. A Flume agent is a (JVM) process that hosts the components through which events flow from an external source to the next destination (hop).



A Flume source consumes events delivered to it by an external source like a web server. The external source sends events to Flume in a format that is recognized by the target Flume source. For example, an Avro Flume source can be used to receive Avro events from Avro clients or other Flume agents in the flow that send events from an Avro sink. A similar flow can be defined using a Thrift Flume Source to receive events from a Thrift Sink or a Flume Thrift Rpc Client or Thrift clients written in any language generated from the Flume thrift protocol.When a Flume source receives an event, it stores it into one or more channels. The channel is a passive store that keeps the event until it's consumed by a Flume sink. The file channel is one example – it is backed by the local filesystem. The sink removes the event from the channel and puts it into an external repository like HDFS (via Flume HDFS sink) or forwards it to the Flume source of the next Flume agent (next hop) in the flow. The source and sink within the given agent run asynchronously with the events staged in the channel.

SQOOP :

Sqoop is a tool designed to transfer data between Hadoop and relational databases. You can use Sqoop to import data from a relational database management system (RDBMS) such as MySQL or Oracle into the Hadoop Distributed File System (HDFS), transform the data in Hadoop MapReduce, and then export the data back into an RDBMS.

Sqoop automates most of this process, relying on the database to describe the schema for the data to be imported. Sqoop uses MapReduce to import and export the data, which provides parallel operation as well as fault tolerance.

Sqoop Tools

HIVE :

Apache **Hive** is a **data** warehouse software project built on top of Apache Hadoop for providing **data** summarization, query and analysis. **Hive** gives an SQL-like interface to query **data** stored in various databases and file systems that integrate with Hadoop.

Apache Hive is a data warehouse software project built on top of Apache Hadoop for providing data summarization, query and analysis.[2] Hive gives an SQL-like interface to query data stored in various databases and file systems that integrate with Hadoop. Traditional SQL queries must be implemented in the MapReduce Java API to execute SQL applications and queries over distributed data. Hive provides the necessary SQL abstraction to integrate SQL-like queries (HiveQL) into the underlying Java without the need to implement queries in the low-level Java API. Since most data warehousing applications work with SQL-based querying languages, Hive aids portability of SQL-based applications to Hadoop

PIG :

Pig is a high level scripting language that is used with Apache Hadoop. **Pig** enables **data** workers to write complex **data** transformations without knowing Java. **Pig's** simple SQL-like scripting language is called **Pig** Latin, and appeals to developers already familiar with scripting languages and SQL.

Apache Pig[1] is a high-level platform for creating programs that run on Apache Hadoop. The language for this platform is called **Pig Latin**.[1] Pig can execute its Hadoop jobs in MapReduce, Apache Tez, or Apache Spark[*citation needed*]. Pig Latin 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. Pig Latin can be extended using user-defined functions (UDFs) which the user can write in Java, Python, JavaScript, Ruby or Groovy