Second phase / Advanced learning:

It is all around the data that too Huge data.

Big Data – Hadoop (MapReduce, YARN, HIVE, HQL, hbase, Scoop, Pig, Zookeeper)

You can learn all the above easily by installing Cloudera VM.

Apache Scala – It’s a programming language (OOP and FP)

Machine learning programs / Messaging Queuing – RabbitMQ / KAFKA

Spark – It’s a data Analytical processing Framework

This Sparking can be done using many languages like R, JAVA, PYTHON and SCALA (Scalable language)

When you write spark things using Python – Pyspark

By default I think it goes with scala.

Cloud computing : PAAS (PLATFORM AS A SERVICE), SAAS (SOFTWARE AS A SERVICE), IAAS (INFRASTRUCTURE AS A SERVICE)

And the below are the key Cloud vendors:

GCP (Google Cloud Platform)

AZURE (provided by Microsoft)  
**AWS** (Amazon Web Services)

You need to deliver a project using python/scala/spark/AWS

Big data – Data Analysis

Big data : Peta bytes

Extremely huge data

Data can be Structured / Unstructured/ Semi Structured

Characteristics of Big Data:

3Vs in Bigdata (Volume, variety, Velocity)

Now they have 2 more

5Vs (Volume, variety, Velocity, Veracity, Value)

Volume – large amount of data

Velocity – the speed / turn over time in which such bigdata are generated

Variety – the type of data

Veracity – to check on the wrong data in tat Big data

Value – to produce useful output

This Bigdata works based on the algorithm called MapReduce

You can do Bigdata things on the Cloudera Virtual machine.

**Big data Frameworks::**

Apache Hadoop : is meant for parallel processing and distributed storage (HDFS – Hadoop Distributed File System, MapReduce)

Apache Spark: Data processing Framework

Apache Kafka: Streaming Framework (AKKA)

Apache Cassandra: Distributed NoSQL DBMS

Hadoop: can be downloaded freely and installed on your local systems.

Hadoop is written in Java.

Cluster – node pattern – Distributed system.

Hadoop – Batch / offline processing.

Several Modules in Hadoop:

HDFS: I will break that huge file into blocks and save it onto the HDFS in the form of the Nodes. Cluster Management.

YARN: Yet Another Resource Negotiator – to deal with several jobs – helps in the scheduling on the cluster.

MAPREDUCE: is a java based program – it will hold the data in the key – value pair.

Again this is to meet the parallel processing.

Map: It will hold the input data in the key-value pair.

Reduce: The output of the Map and combines the broken elements and gives desired output.

HADOOP COMMON: Is nothing but the common Hadoop java libraries to perform Hadoop things.

Hadoop – Master slave architecture.

NameNode: Master – only one master allowed. Lets you open the file, close file.

DataNode: several datanodes and are slaves.

All the data blocks which were broken are in the datanodes. It lets you writing-read-edit the files.

There are several commands w.r.t to HDFS system.

YARN : Components of YARN: Job scheduler

Client: who submits the MapReduce program/job

Resource manager: to manage the resources among many clusters

Node Manager: to manage several nodes.

Application Master: it monitors the jobs running.

Map Reduce – is a java program / tool and has 2 phases:

Mapper phase:

Reducer phase:

Sort and shuffle:

Hadoop Ecosystem:

1. HDFS : primary storage component. Many command prompt type of commands are der.
2. MapReduce: It can process peta bytes of data and also takes care of failures
3. Yarn: Job scheduler
4. Hive: Datawarehouse system – OLAP : Summarizes the data, queries the data and analyses the data.

HQL – HIVE Query Language. It implicitly translates the HQL into the MapReduce Jobs which is what Hadoop wants for execution.

1. Hbase: It is a distributed database system on HDFS. Hbase is a structured DB and is a NoSQL.
2. Pig: It takes anything on the Hadoop system. It is a language platform – uses pigLatin language.
3. Sqoop: It is for export – import of data between HDFS and external sources. (Oracle, Mysql, flatfile, MongoDB)
4. Zookeeper: is a security guard for Hadoop ecosystem, and manages the several clusters on the HDFS. Tracks all the transcations happening.
5. Oozie: workflow scheduler – He can combine many such Hadoop jobs in a sequence as one logical unit.
6. Hcatalog: It is a storage management.
7. Avro
8. Thrift
9. Mahout: can create ML algorithms – to find out the patterns or any ML activity.
10. Flume
11. Ambari