Big Data

Monthly Newsletter - Gopal Mukkamala 2.26.2012 Issue 1

Inside

Learning

Cloudera
HPCC
IBM

Best of all:-
Download and learn yourself.

Technology Mix

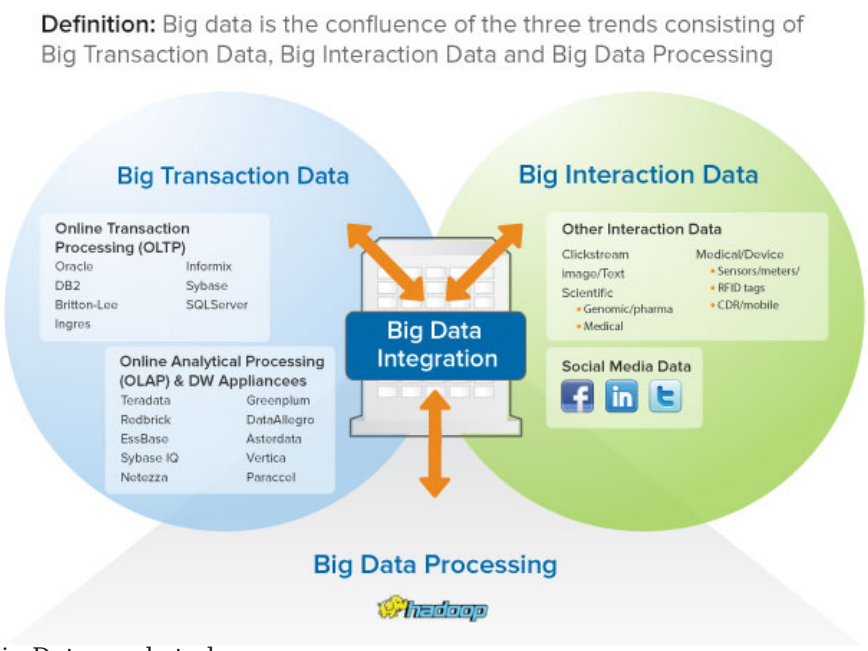
Hadoop/Hbase/Pig

The tecnology mix for Big Data includes but is not limited to :-
Hadoop - the cluster
Hbase - the database
Hive - Sql Like interface
Pig - Mining

Where to Start

Download

Apache project has all of the downloads freely available. The challenge is in making different versions to work together.



Big Data market place.

In a nutshell, Big Data is processing vast amounts of data. The beauty of it lies in the fact that you can process it on Commodity servers.

Hadoop is built from ground up for fault tolerance which means that the software is aware that a hardware component within its cluster will fail.

It is as fairly easy to remove a machine from the cluster as it is to add it back.

If you have worked on Oracle RAC technology you can appreaciate this effort by apache community on how difficult it is for interprocess communication where split brain scenarios, node reboot etc, are part of daily life.

The part that makes big data very useful is the map/reduce algorithms.

Map reduce takes a key/value pair further reduces into another key value pair sounds easy. Well you really have to dwelve into it to understand and appreciate it.

Last but not least, Big Data is here to Stay. It will be worth your investment to keep yourself posted on this fantastic technology. - Good Luck.

Learning big data is not hard, however be ready to spend quiet a bit of time on various technologies. It all starts with Linux(Debian,Ubuntu,RedHat etc.,), once you have a stable system up and running you can start your journey by downloading :-

1.**Oracle JDK 1.6, Hadoop.**
Hadoop version 0.20.x is the current stable version.
Download the software in your system once it is installed you can run java -version java version "1.6.0_26"

2. Download Hadoop from http://hadoop.apache.org/ extract it to a location in your system and set the HADOOP_HOME environment variable. At a minimum you have to set the JAVA_HOME variable in the file hadoop-env.sh
export JAVA_HOME=/usr/lib/jvm/java-6-sun/jre

3. Set the value in core-site.xml
hadoop.tmp.dir
/home/oracle/myhadoop/tmp...

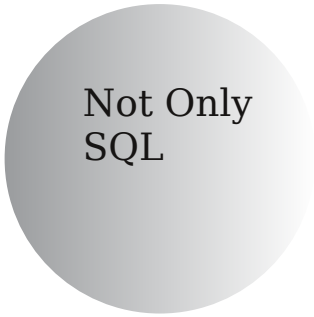
fs.default.name
hdfs://localhost:54310
The port 54310 is where your HDFS is listening on.

Verifying the runtime

In order to verify the run time issue the following:-
1783 NameNode
23241 Jps
1868 DataNode
2131 TaskTracker
2280 HMaster
2039 JobTracker
1966 SecondaryNameNode

netstat -a |grep -i 5431 you should see your ports listed there.

My Hadoop system is running in psuedo distributed mode which means we are running on a single system.



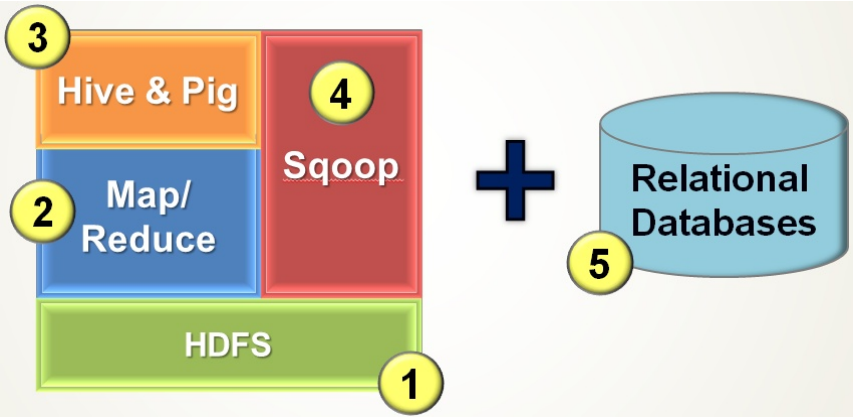
Map/Reduce Program

Map/Reduce java program in apache site is outdated. You would require a latest version of the program in order to compile and run it.

http://myexadata.blogspot.com/p/hadoop.html

Once you get a feel of it, it is fairly easy to run it.

In the market place MR itself will determine who the leader will be based on the use case scenarios they can build on MR.



HBase

HBase is a great software which tightly integrates with Hadoop HDFS.
Start HBase:-
Similar to starting hadoop you can start hbase by running $HBASE_HOME/bin/start-hbase.sh

Hbase Shell Example:-
Create 'hospital','patient';
Put 'hospital','1','patient:name','Gopal'
put 'hospital','1,'patient:diagnosis','Healthy'

The format of the data to put and get is :-
put <table><row><columnfamily><value>
similarly you can retrieve the rows by running
get 'hospital','1'
the output of the command will be :-

```
hbase(main):005:0> get 'hospital','1'
COLUMN                CELL
 patient:diagnosis    timestamp=1330262704788, value=Healthy
 patient:name         timestamp=1330262525035, value=Gopal
2 row(s) in 0.0560 seconds
```

The beauty of HBase is in the fact that it leverages Hadoop's HDFS capabilities in terms of data replication.
3 Copies of every block are made, this is due to the fact that at any moment of time you can afford to loose two machines at the same time without any outage.

Hive/Confluence/SQoop

Similar to Hbase You can download Hive. Hive has been renamed to confluence. Hive has a added advantage of running your Queries in something called HiveQL.

What is Sqoop
Sqoop (“SQL-to-Hadoop”) is a straightforward command-line tool with the following capabilities:
Imports individual tables or entire databases to files in HDFS
Generates Java classes to allow you to interact with your imported data
Provides the ability to import from SQL databases straight into your Hive data warehouse.

Conclusion

Big Data is here to stay. As we go along in this journey the major players will strive to make the loading, reading and updating of data as seamless as possible.

Hadoop Distributed File System (HDFS) – objectives are load balancing, fact access and fault tolerance, designed with the expectations that hardware/software failures are a fact of life.

MapReduce – framework for writing/executing distributed, fault tolerant algorithms functions map which divided a large problem into smaller problems and then performs the same function on all smaller problems and reduce which then combines the results.

Hive & Pig – Hive was created by Facebook and is SQL-like, while Pig was created by Yahoo and is more procedural; both target MapReduce jobs. However due to the complexity of MapReduce, HiveQL was created to combine the best features of SQL with MapReduce
Sqoop – package for moving data between HDFS and relational DB systems via command line load and unload utilities

Gopal Mukkamala

Email:-
mhgopal@gmail.com

I love Big Data. It is just amazing on how these technologies have come together and evolved.

Imagine you running your processes without any performance issues on Terrabytes of data.