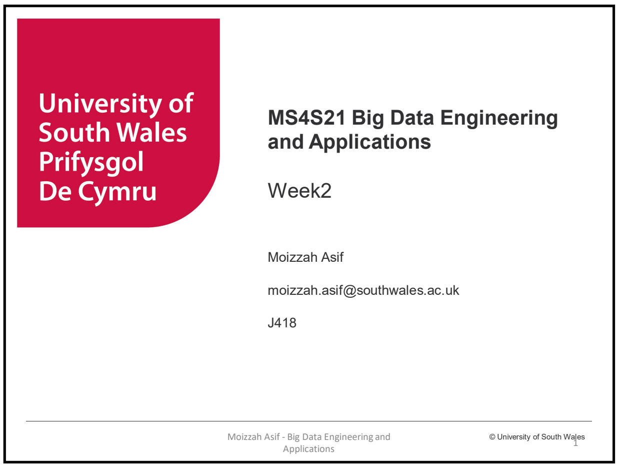


University of South Wales
Prifysgol De Cymru

MS4S21 Big Data Engineering and Applications

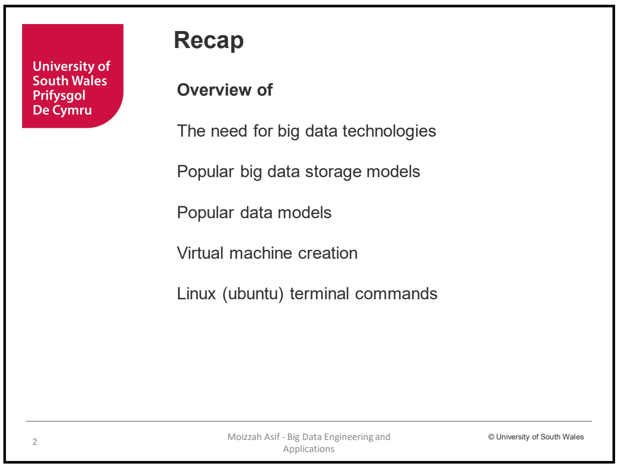
Week2

Moizzah Asif

moizzah.asif@southwales.ac.uk

J418

1



University of South Wales
Prifysgol De Cymru

Recap

Overview of

The need for big data technologies

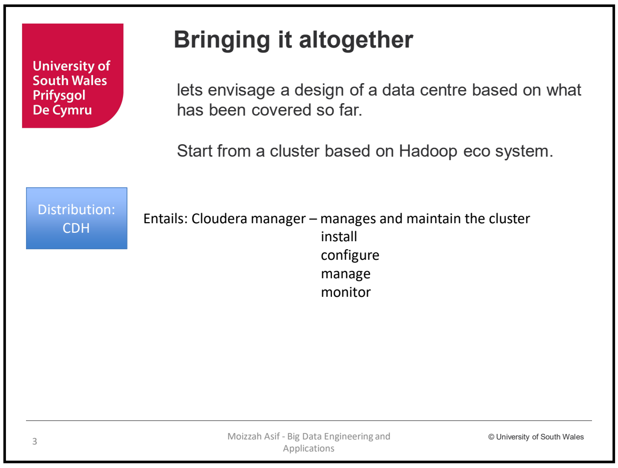
Popular big data storage models

Popular data models

Virtual machine creation

Linux (ubuntu) terminal commands

2



University of South Wales
Prifysgol De Cymru

Bringing it altogether

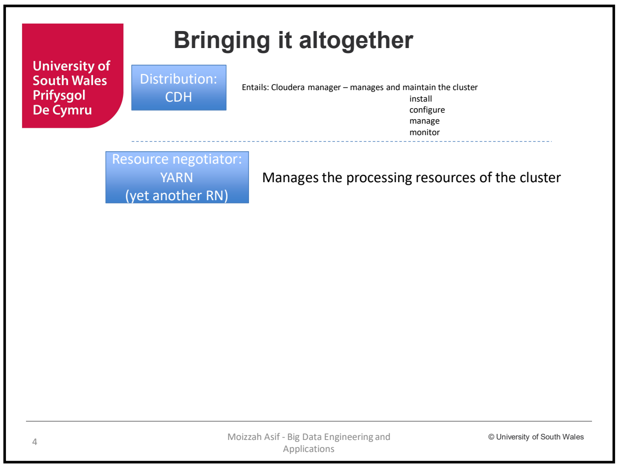
lets envisage a design of a data centre based on what has been covered so far.

Start from a cluster based on Hadoop eco system.

Distribution: CDH

Entails: Cloudera manager – manages and maintain the cluster
install
configure
manage
monitor

3



Bringing it altogether

University of South Wales Prifysgol De Cymru

Distribution: CDH

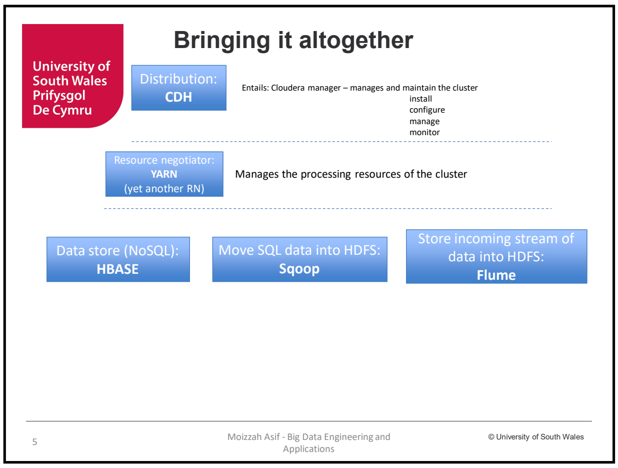
Entails: Cloudera manager – manages and maintain the cluster
install
configure
manage
monitor

Resource negotiator: YARN (yet another RN)

Manages the processing resources of the cluster

4 Moizzah Asif - Big Data Engineering and Applications © University of South Wales

4



Bringing it altogether

University of South Wales Prifysgol De Cymru

Distribution: **CDH**

Entails: Cloudera manager – manages and maintain the cluster
install
configure
manage
monitor

Resource negotiator: **YARN** (yet another RN)

Manages the processing resources of the cluster

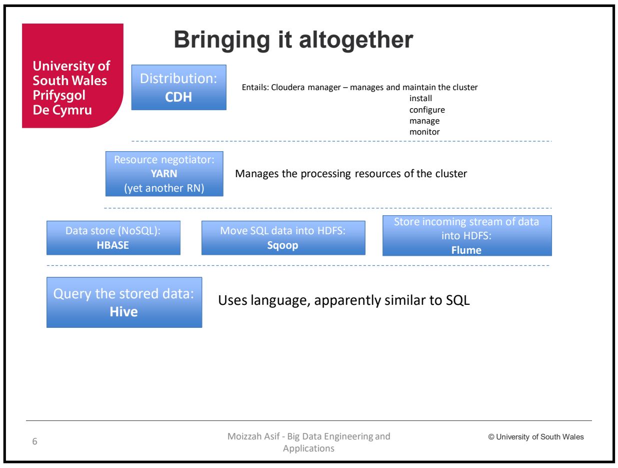
Data store (NoSQL): **HBASE**

Move SQL data into HDFS: **Sqoop**

Store incoming stream of data into HDFS: **Flume**

5 Moizzah Asif - Big Data Engineering and Applications © University of South Wales

5



Bringing it altogether

University of South Wales Prifysgol De Cymru

Distribution: **CDH**

Entails: Cloudera manager – manages and maintain the cluster
install
configure
manage
monitor

Resource negotiator: **YARN** (yet another RN)

Manages the processing resources of the cluster

Data store (NoSQL): **HBASE**

Move SQL data into HDFS: **Sqoop**

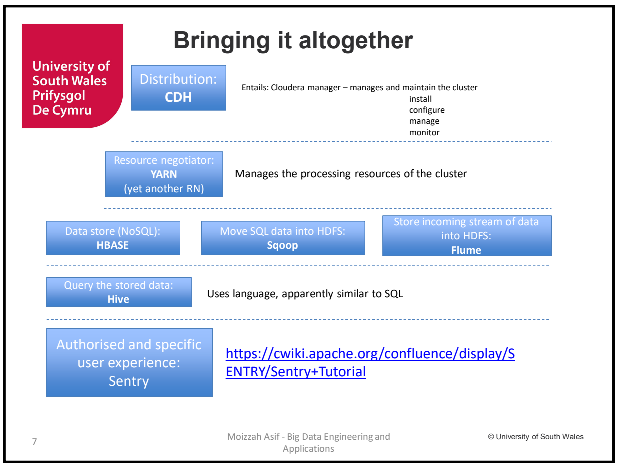
Store incoming stream of data into HDFS: **Flume**

Query the stored data: **Hive**

Uses language, apparently similar to SQL

6 Moizzah Asif - Big Data Engineering and Applications © University of South Wales

6



University of South Wales Prifysgol De Cymru

Bringing it altogether

Distribution: **CDH**

Entails: Cloudera manager – manages and maintain the cluster
install
configure
manage
monitor

Resource negotiator: **YARN** (yet another RN)

Manages the processing resources of the cluster

Data store (NoSQL): **HBASE**

Move SQL data into HDFS: **Sqoop**

Store incoming stream of data into HDFS: **Flume**

Query the stored data: **Hive**

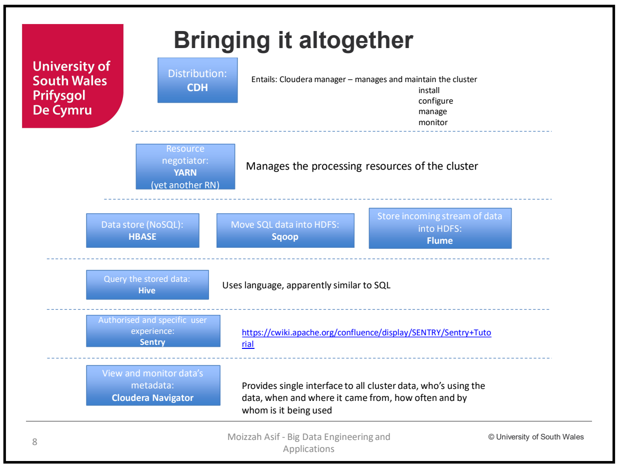
Uses language, apparently similar to SQL

Authorised and specific user experience: Sentry

https://cwiki.apache.org/confluence/display/SENTRY/Sentry+Tutorial

7 Moizzah Asif - Big Data Engineering and Applications © University of South Wales

7



University of South Wales Prifysgol De Cymru

Bringing it altogether

Distribution: **CDH**

Entails: Cloudera manager – manages and maintain the cluster
install
configure
manage
monitor

Resource negotiator: **YARN** (yet another RN)

Manages the processing resources of the cluster

Data store (NoSQL): **HBASE**

Move SQL data into HDFS: **Sqoop**

Store incoming stream of data into HDFS: **Flume**

Query the stored data: **Hive**

Uses language, apparently similar to SQL

Authorised and specific user experience: **Sentry**

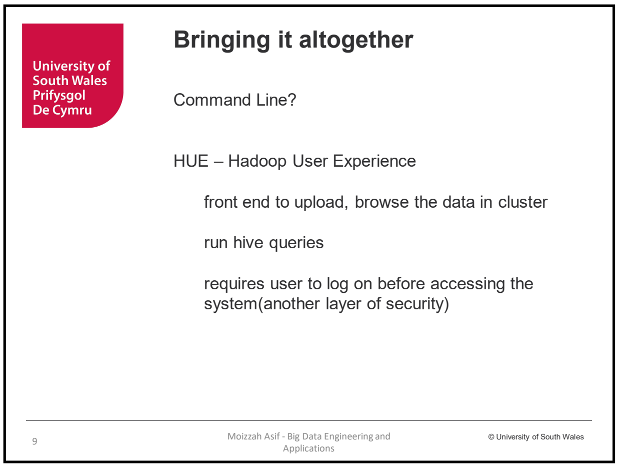
https://cwiki.apache.org/confluence/display/SENTRY/Sentry+Tutorial

View and monitor data's metadata: **Cloudera Navigator**

Provides single interface to all cluster data, who's using the data, when and where it came from, how often and by whom is it being used

8 Moizzah Asif - Big Data Engineering and Applications © University of South Wales

8



University of South Wales Prifysgol De Cymru

Bringing it altogether

Command Line?

HUE – Hadoop User Experience

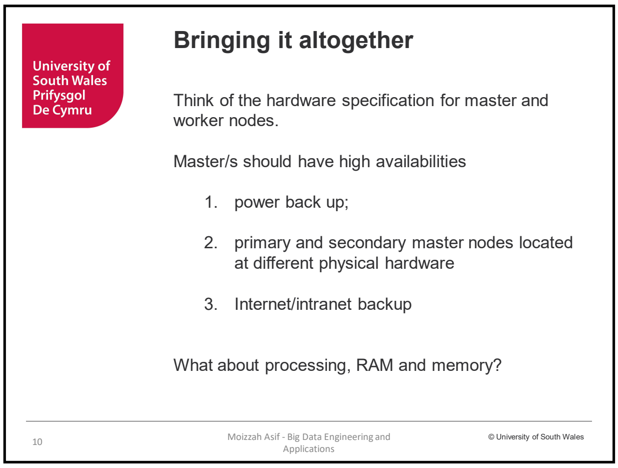
front end to upload, browse the data in cluster

run hive queries

requires user to log on before accessing the system(another layer of security)

9 Moizzah Asif - Big Data Engineering and Applications © University of South Wales

9



Bringing it altogether

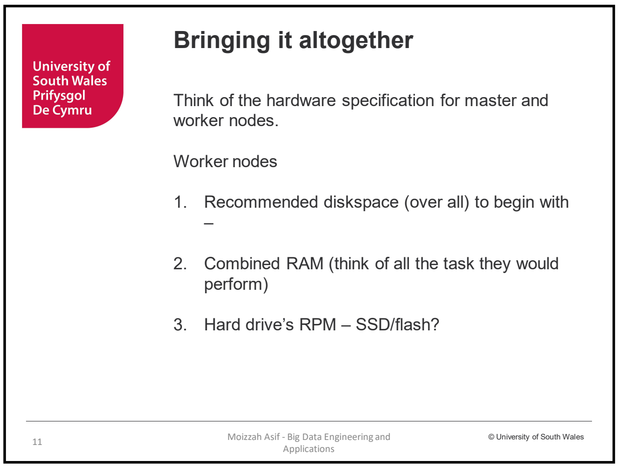
Think of the hardware specification for master and worker nodes.

Master/s should have high availabilities

1. power back up;
2. primary and secondary master nodes located at different physical hardware
3. Internet/intranet backup

What about processing, RAM and memory?

10



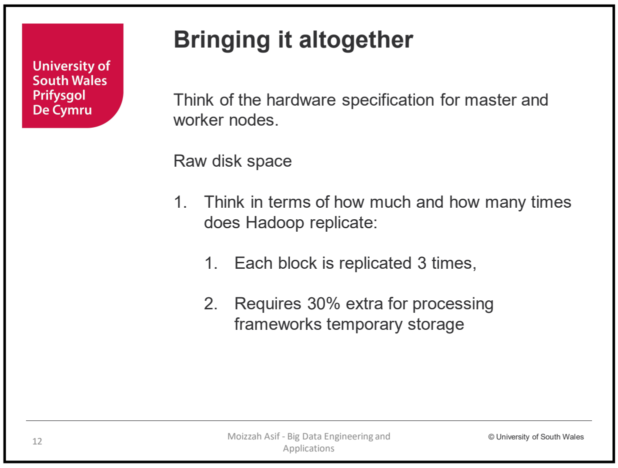
Bringing it altogether

Think of the hardware specification for master and worker nodes.

Worker nodes

1. Recommended diskspace (over all) to begin with –
2. Combined RAM (think of all the task they would perform)
3. Hard drive's RPM – SSD/flash?

11



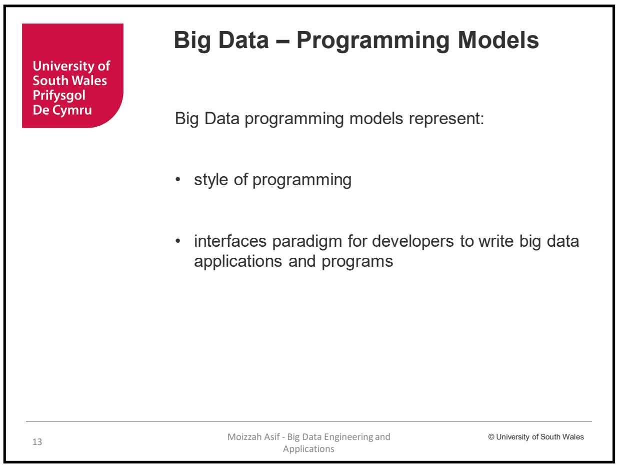
Bringing it altogether

Think of the hardware specification for master and worker nodes.

Raw disk space

1. Think in terms of how much and how many times does Hadoop replicate:
   1. Each block is replicated 3 times,
   2. Requires 30% extra for processing frameworks temporary storage

12



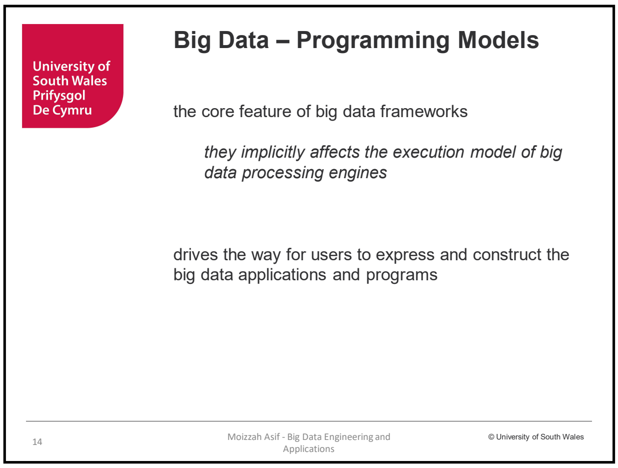
Big Data – Programming Models

Big Data programming models represent:

- style of programming
- interfaces paradigm for developers to write big data applications and programs

13 Moizzah Asif - Big Data Engineering and Applications © University of South Wales

13



Big Data – Programming Models

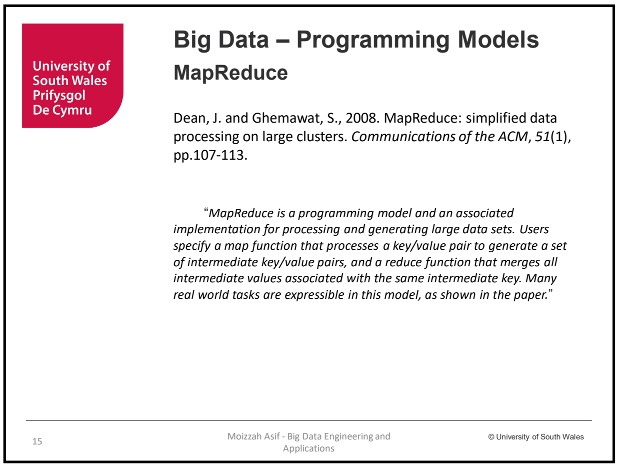
the core feature of big data frameworks

they implicitly affects the execution model of big data processing engines

drives the way for users to express and construct the big data applications and programs

14 Moizzah Asif - Big Data Engineering and Applications © University of South Wales

14



Big Data – Programming Models

MapReduce

Dean, J. and Ghemawat, S., 2008. MapReduce: simplified data processing on large clusters. *Communications of the ACM*, *51*(1), pp.107-113.

"MapReduce is a programming model and an associated implementation for processing and generating large data sets. Users specify a map function that processes a key/value pair to generate a set of intermediate key/value pairs, and a reduce function that merges all intermediate values associated with the same intermediate key. Many real world tasks are expressible in this model, as shown in the paper."

15 Moizzah Asif - Big Data Engineering and Applications © University of South Wales

15