# BIG DATA

## Hadoop Ecosystem

**K V Subramaniam**
Computer Science and Engineering

# Hadoop Ecosystem Overview

PES UNIVERSITY ONLINE

## What we have learnt so far..

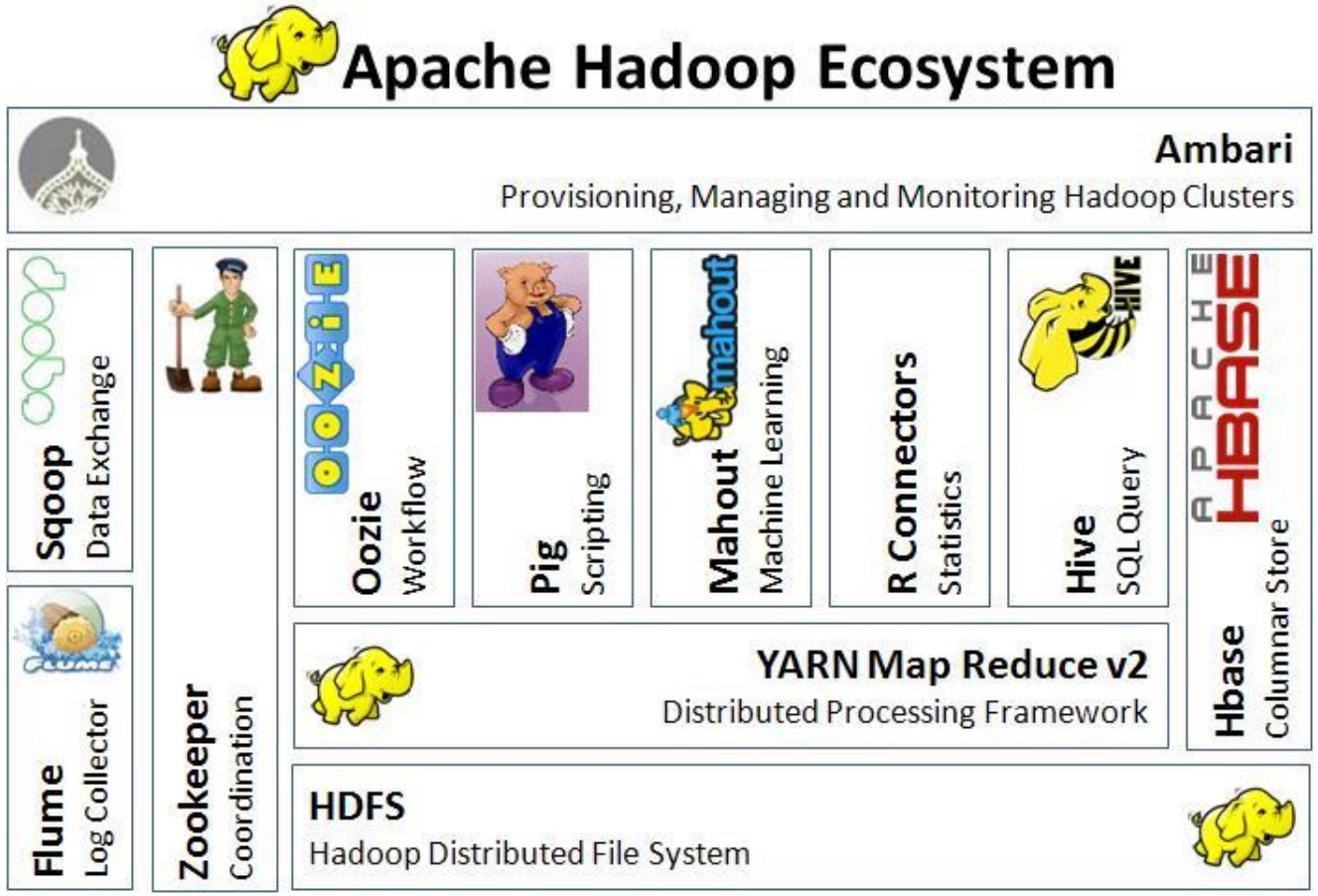
HDFS for storage

And MapReduce (Hadoop) for computation

So where does it all fit in the bigger scheme of a Big Data architecture

# BIG DATA

## Hadoop Ecosystem




www.facebook.com/hadoopers

# Hadoop Workflows: Oozie

## Oozie: Motivational exercise

Suppose we want to build a recommendation system like in AMAZON

What inputs would we need?
How often would we need to update the recommendations (every hour? Every day? EVERY week?)
Is it enough to just build a recommendation algorithm?

that can work PROVIDED the right inputs are given to it.

# BIG DATA

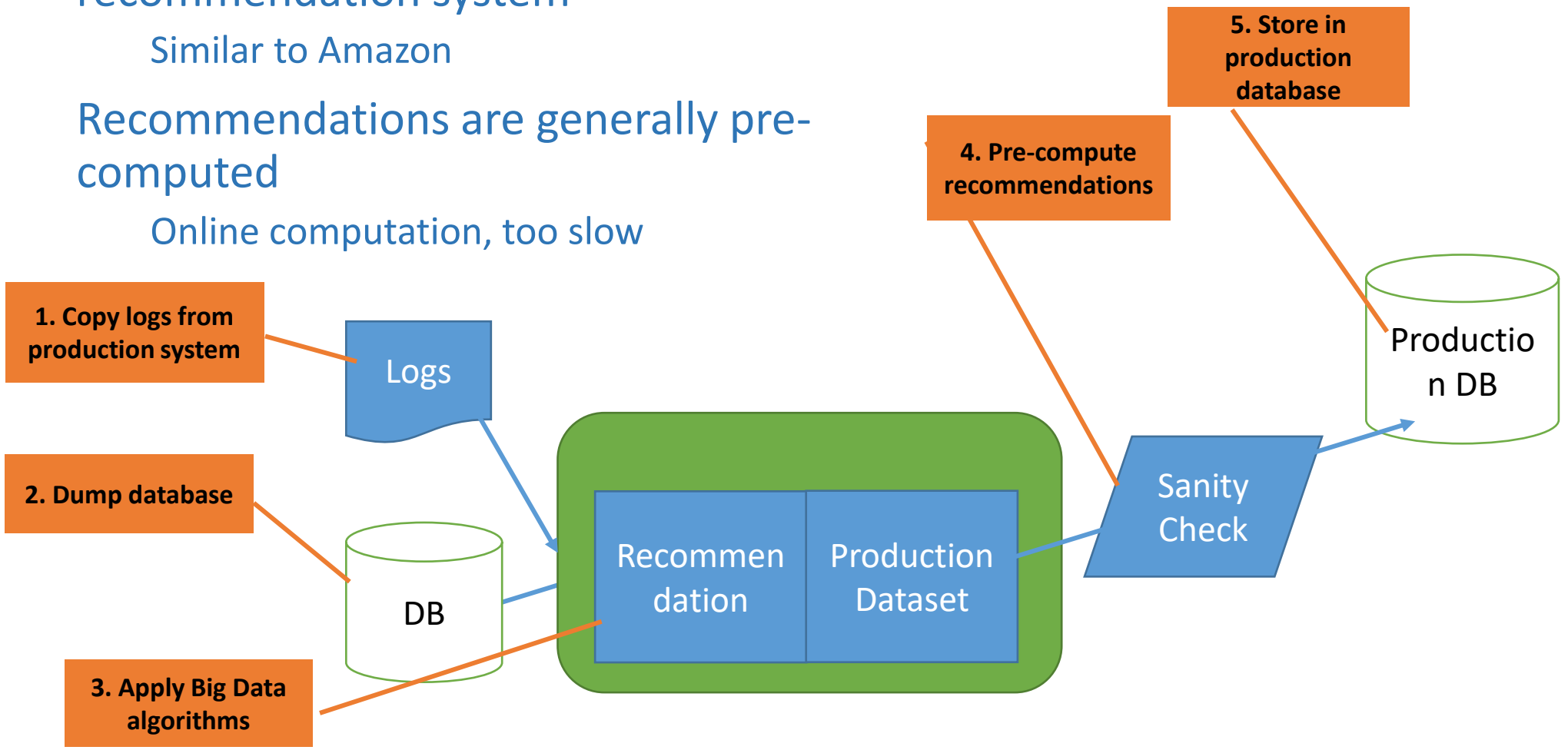
## Hadoop Ecosystem

We want to use the sales to build a recommendation system

- Similar to Amazon

Recommendations are generally pre-computed

- Online computation, too slow



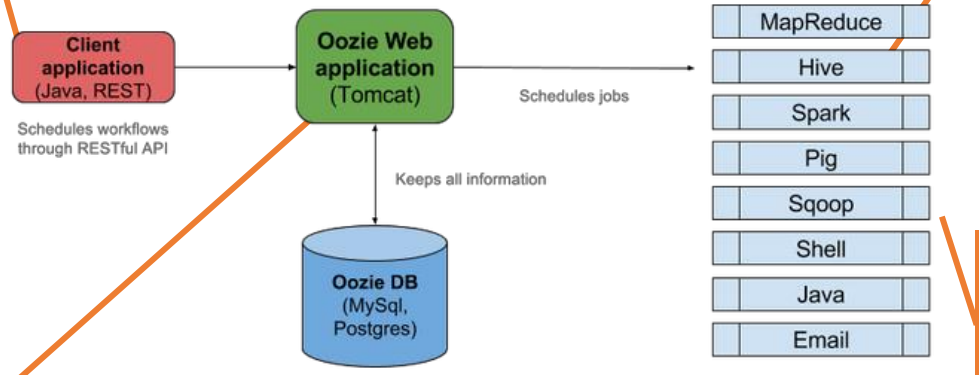
## Workflow Definition

- The sequence of steps is called a *Workflow*
- Workflows are common in data centers
- Users would like to
  - Specify the steps
  - Specify when the steps are to be run
    - Maybe periodically
  - Run the Workflow
  - What to do in case of error

# BIG DATA

## Oozie Architecture

1. User submits Workflow (XML)




3. Some of the applications supported by Oozie

2. Oozie: Tomcat application

4. Workflow can include any of these applications

[Image courtesy: https://oyermolenko.blog/2017/10/01/scheduling-jobs-in-hadoop-through-oozie/](https://oyermolenko.blog/2017/10/01/scheduling-jobs-in-hadoop-through-oozie/)

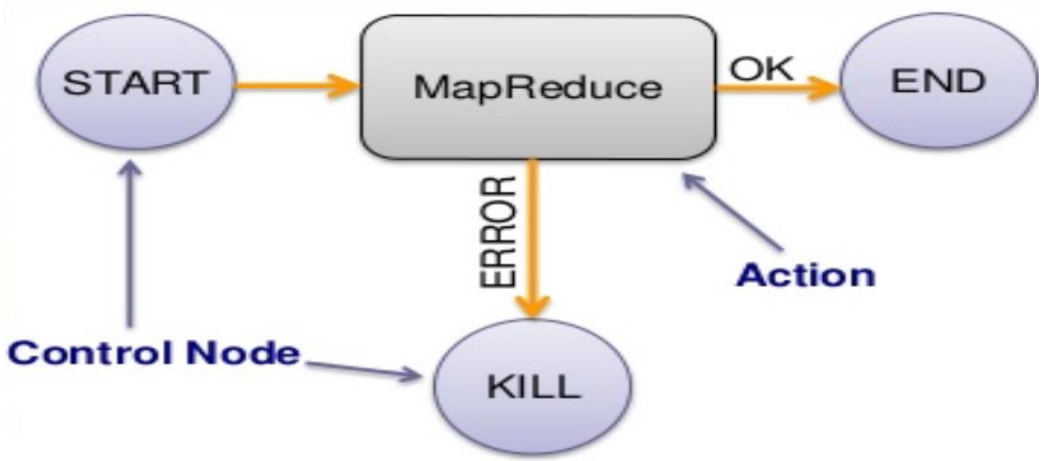
# Oozie Workflow (pictorial)

## Action nodes

- Does something, e.g., run Mapreduce
- Every action node has a normal exit and an error exit

## Control nodes

- Start is the beginning of a Workflow
- End and kill are the end
- Other control nodes
  - Fork (start parallel tasks) and Join (merge parallel tasks)
  - Decision: like switch




DAG Expressing Workflow

# BIG DATA

## Oozie References

T1 Chapter 2.6.1.2

T2 Chapter 7.5

http://oozie.apache.org/docs/4.3.0/index.html

Official Oozie Homepage

https://oyermolenko.blog/2017/10/01/scheduling-jobs-in-hadoop-through-oozie/

A very good introduction to Oozie

# Hadoop Workflows: Ambari

# Ambari : Deploy and Manage Hadoop Clusters

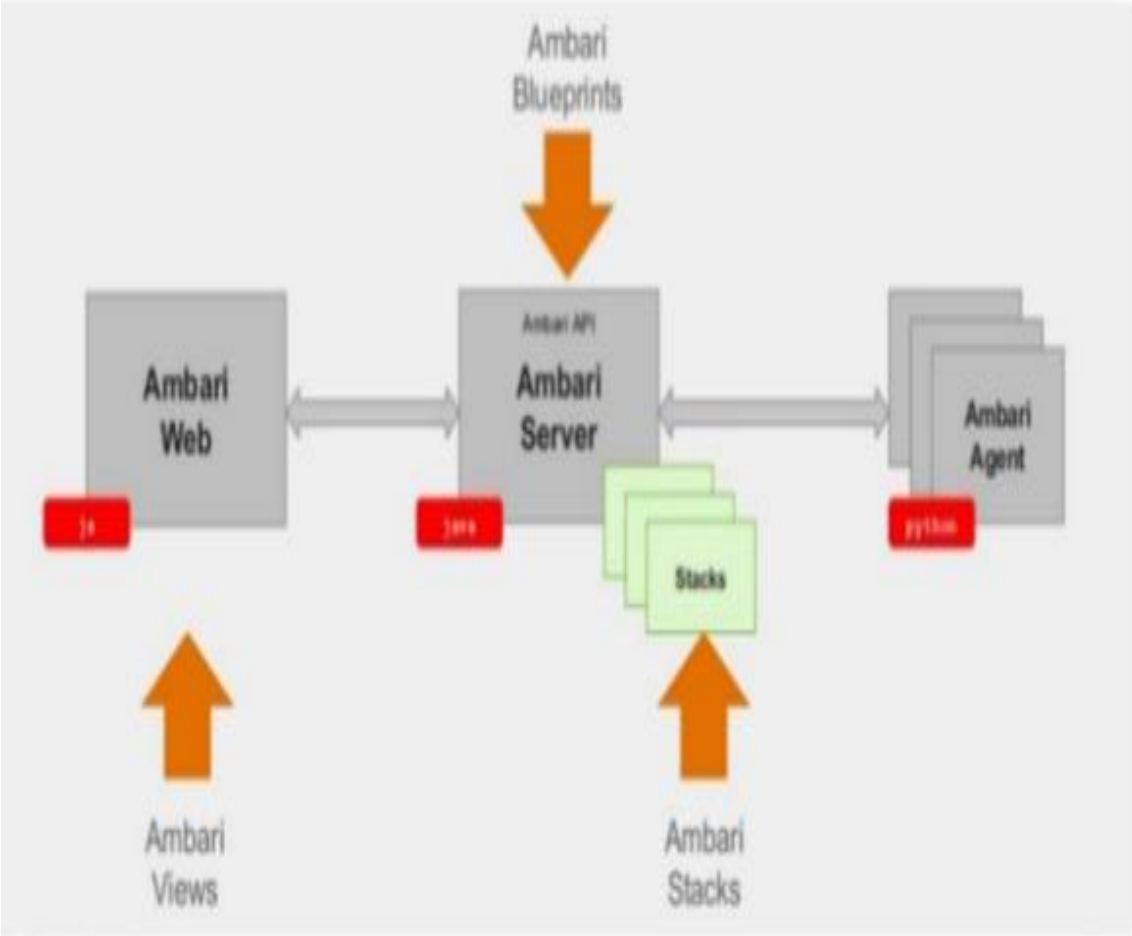
Simplifies Installation, Configuration and Management

Easy, efficient, repeatable creation of clusters

Manages and Monitors clustering

# BIG DATA

## Ambari Architecture




Source: https://www.slideshare.net/hortonworks/managing-enterprise-hadoop-clusters-with-apache-ambari

**Ambari Stacks**

Describes the applications to be installed, eg, Hadoop, its components and its structure

**Ambari Blueprints**

Creation of the cluster

**Ambari Views**

User interface

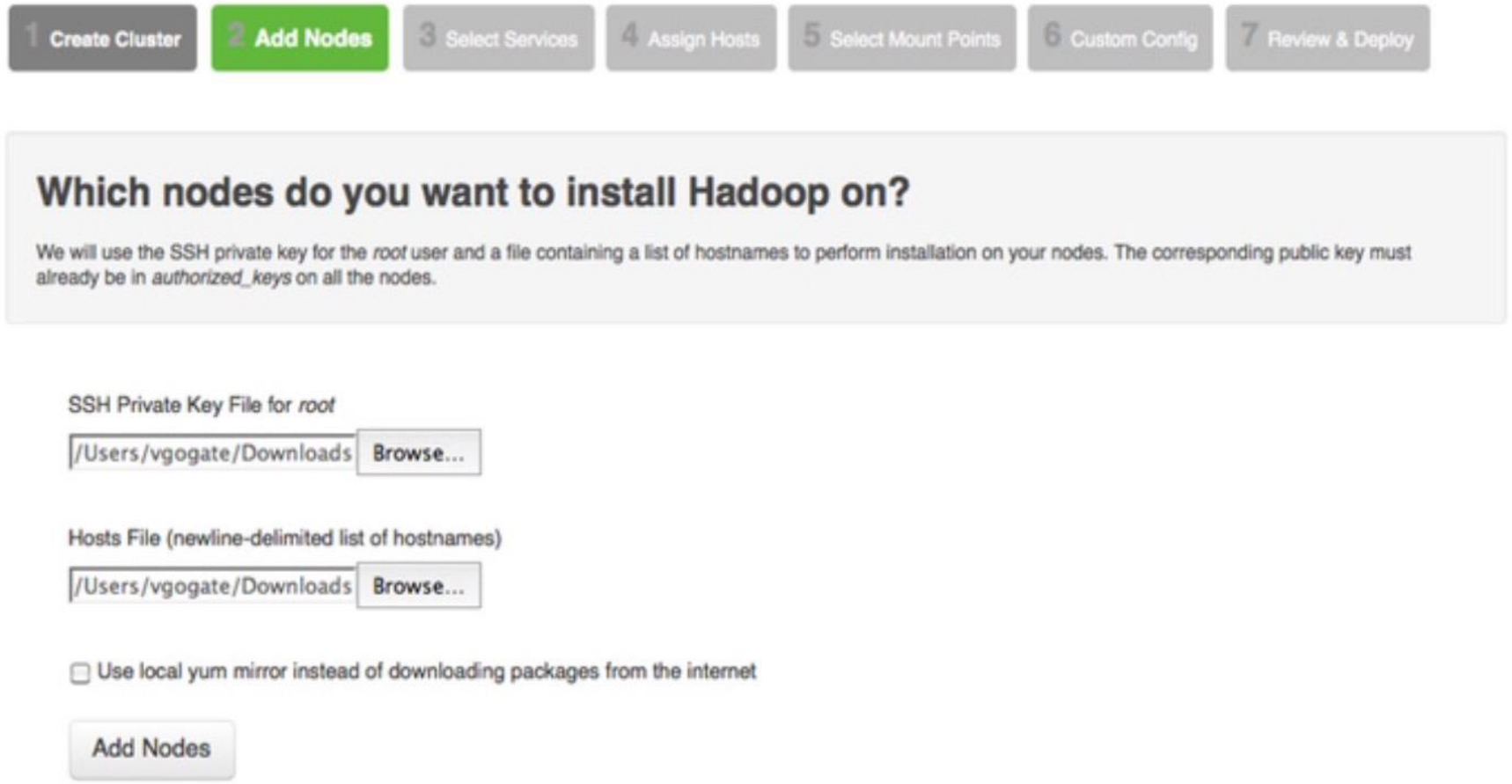
# Ambari Stacks

## What do we need to define in Ambari to install a cluster ?

| Term | Meaning |
|---|---|
| Stack | Set of services, where to get the software packages, e.g. HDP (Hortonworks Data Platform) |
| Service | Components that make up the service e.g, HDFS |
| Component | Building blocks of the service Namenode, Datanode |
| Category | Master, slave, client |

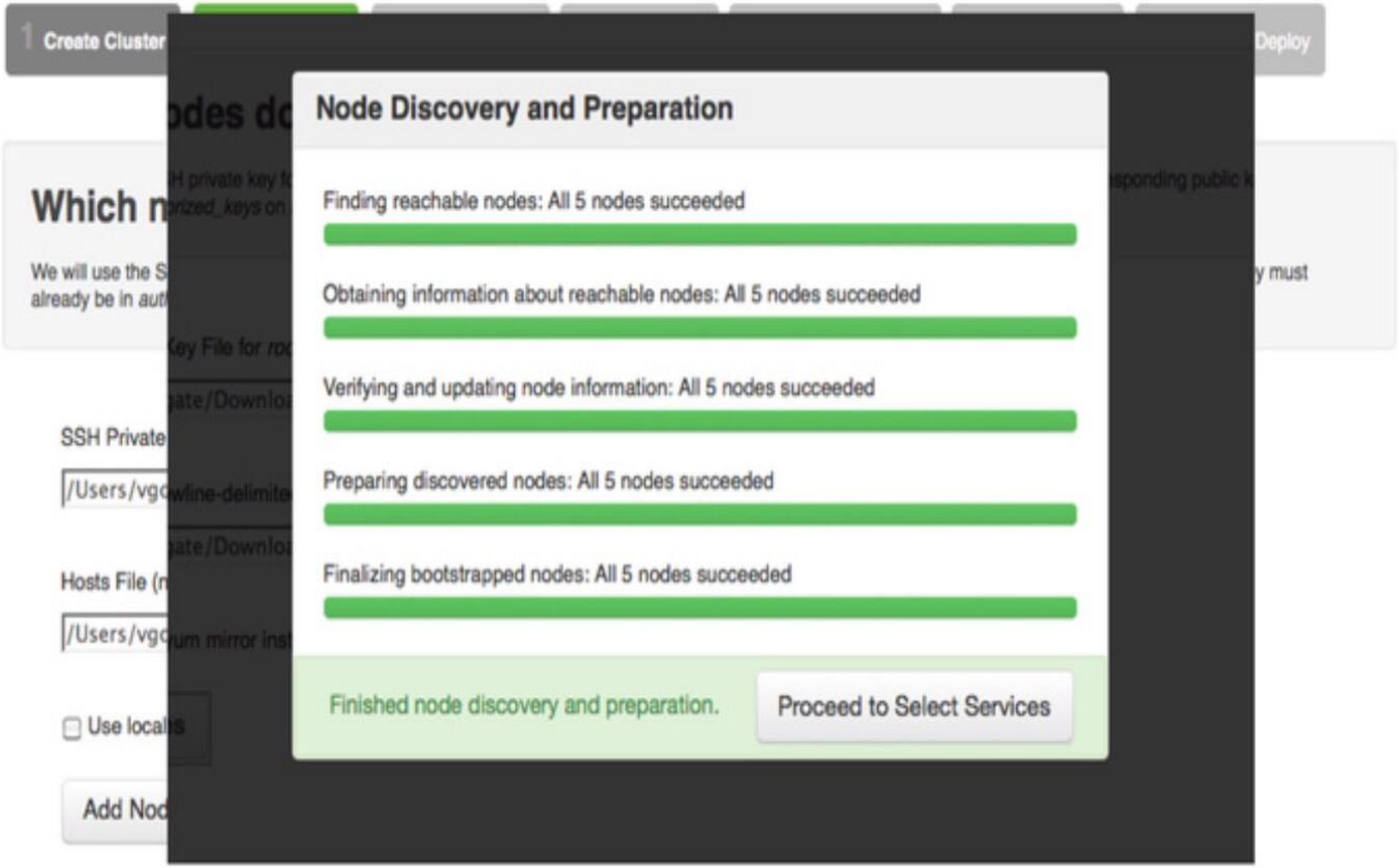
# BIG DATA

## Hadoop Ecosystem



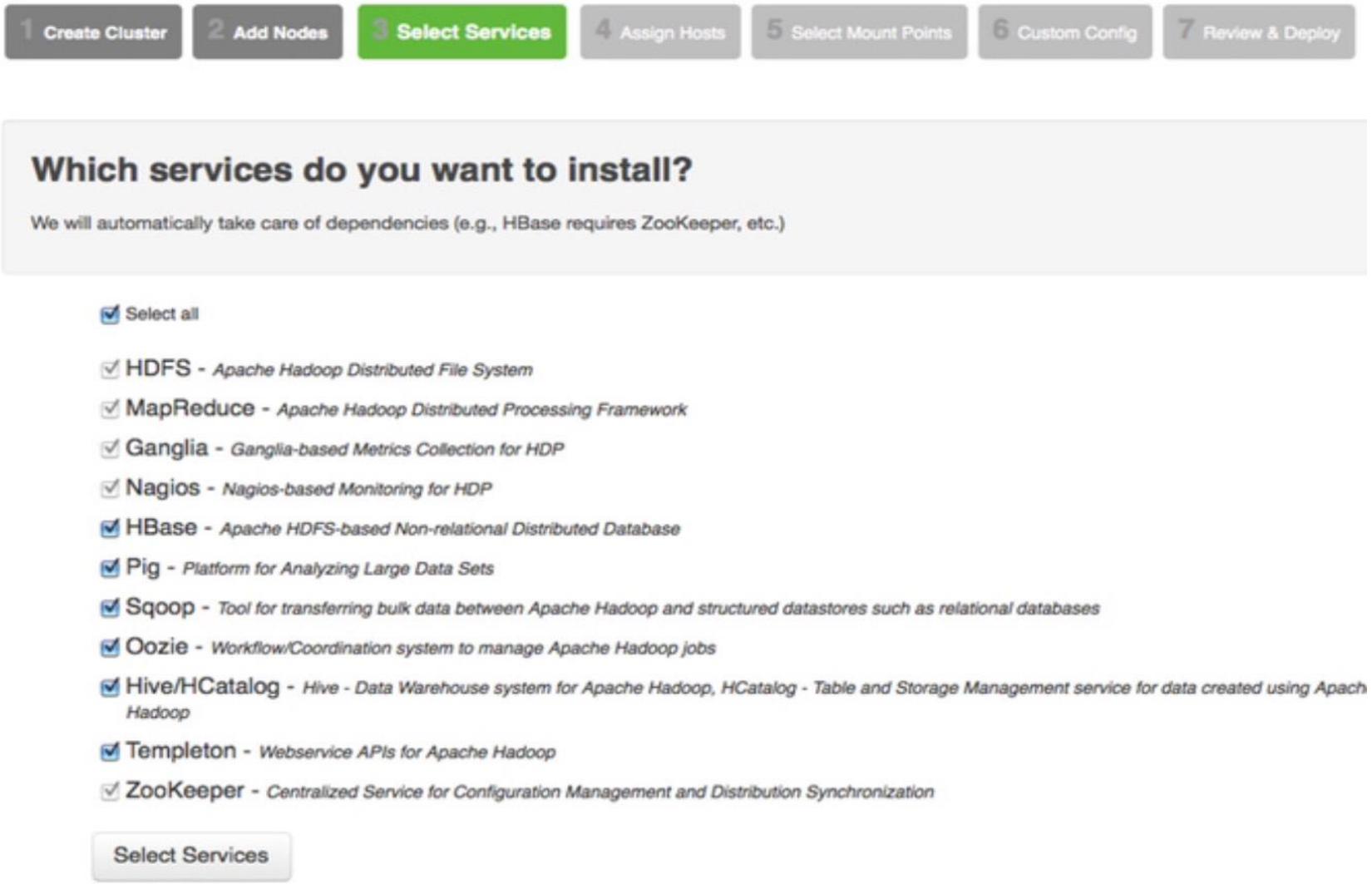
# BIG DATA

## Hadoop Ecosystem



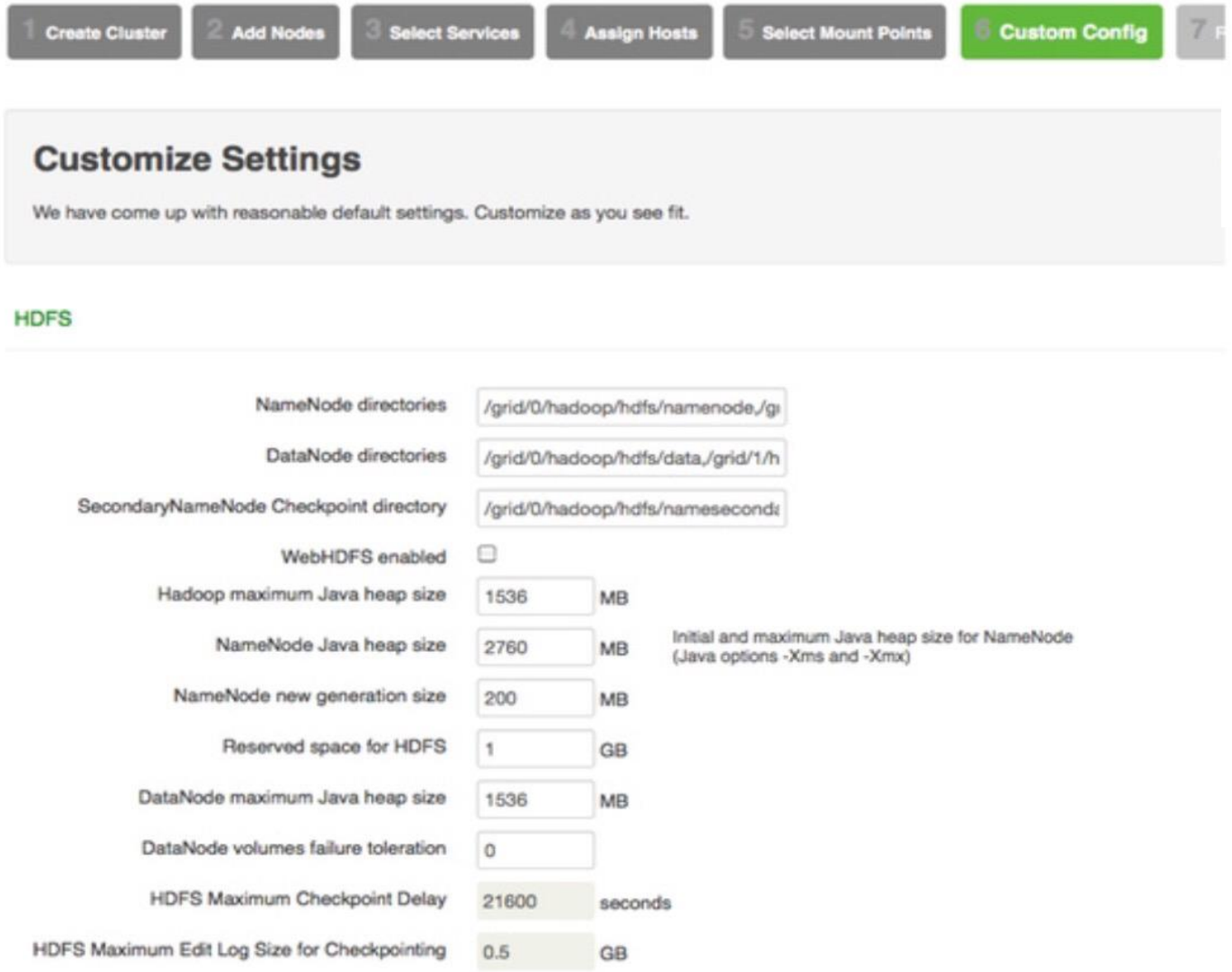
# BIG DATA

## Hadoop Ecosystem



# BIG DATA

## Hadoop Ecosystem



# BIG DATA

## Hadoop Ecosystem

# Pig : Building High-Level Dataflows over Map Reduce

## Disadvantages of MapReduce

For Data analysis, Map-Reduce is too low-level

Writing Map and Reduce code requires retraining.

Something SQL-Like may be better

- HIVE is an option (which we will look at later)
- But how about a scripting language

# BIG DATA

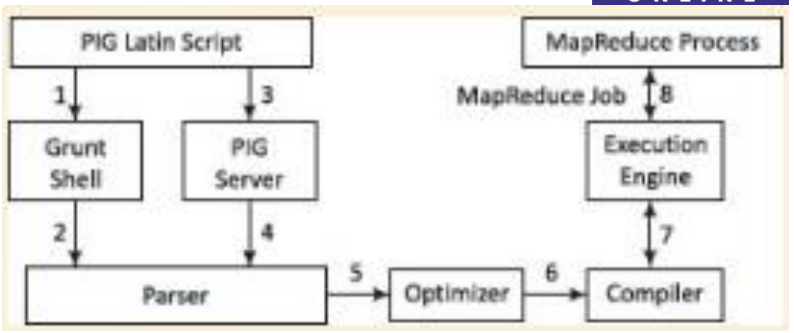
## PIG Introduction

Complex data transformations using scripts

Builtin operators

Interactive shell Grunt

Language Pig Latin




Scripts are internally converted to Map Reduce jobs.

Created @Yahoo

Image courtesy: Big Data Analytics, Rajkamal, Preeti Saxena

# BIG DATA

## Example Data Analysis Task

Find the top 10 most popular IPL matches in each venue.

### Visits

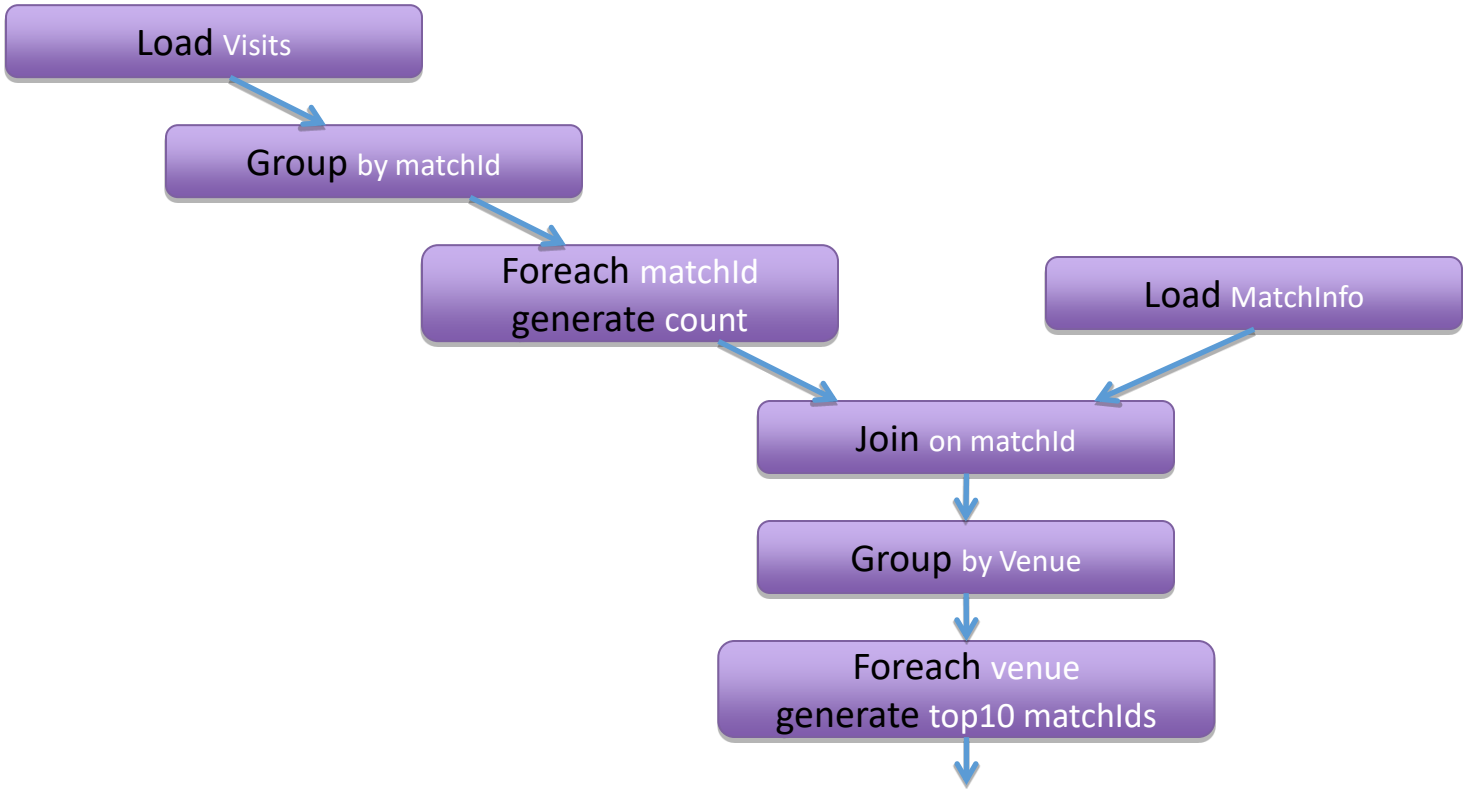
| User | matchId | Time |
|---|---|---|
| RajniKan | Match 400 | 2:00 |
| RajniKan | Match 201 | 5:00 |
| Superman | Match 42 | 10:05 |
| Spiderman | Match 108 | 13:03 |

### MatchInfo

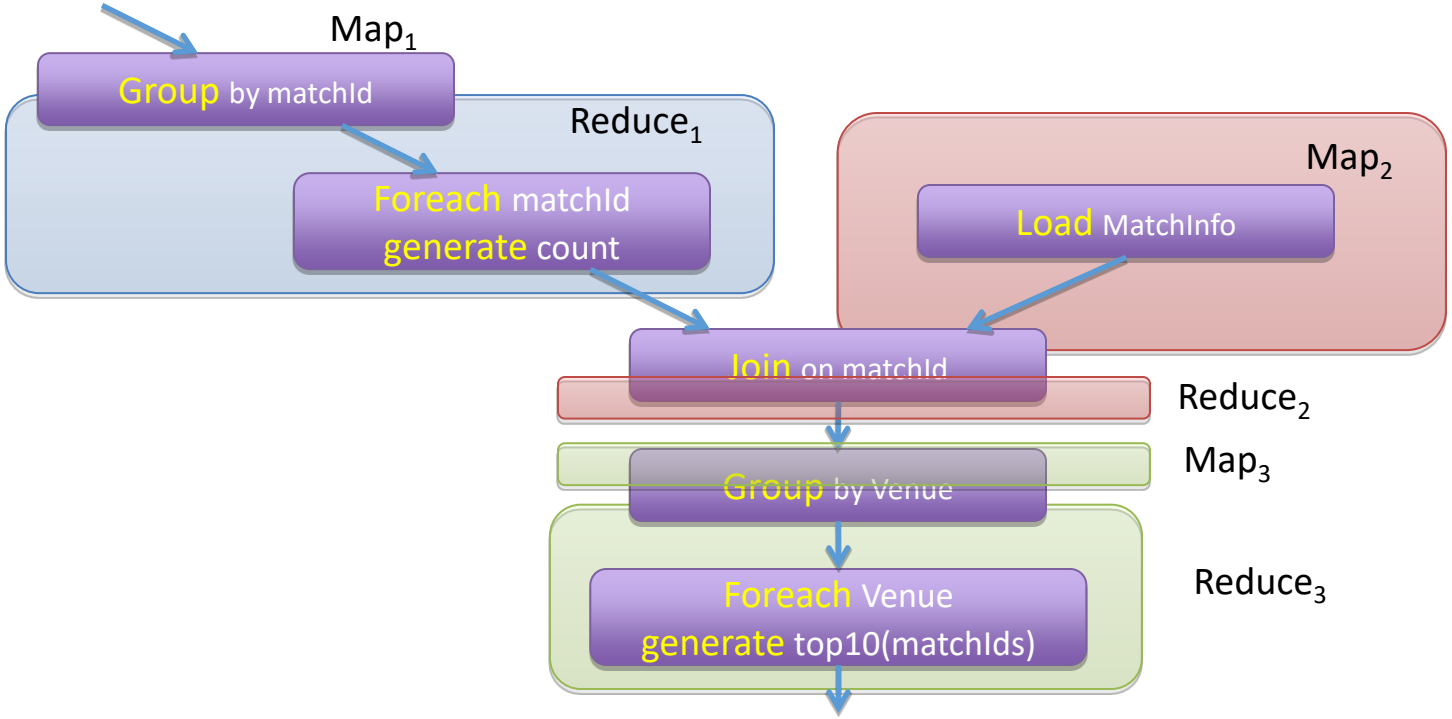
| matchId | Venue | Winner |
|---|---|---|
| Match 108 | Chennai | CSK |
| Match 201 | Bengaluru | RCB |
| Match 42 | Kolkata | DC |
| Match 400 | Mumbai | RCB |

# BIG DATA

## Data Flow



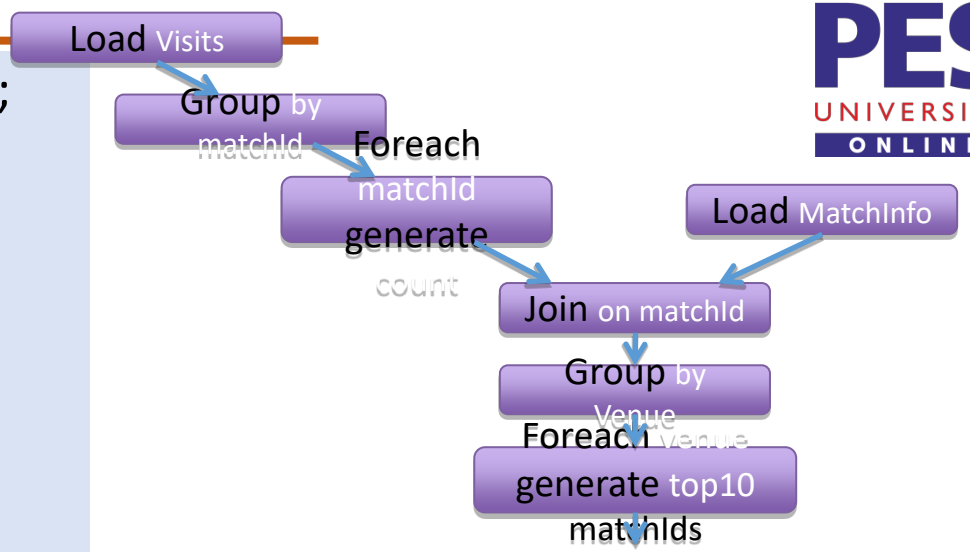
## Compilation into Map-Reduce



# BIG DATA

## In Pig Latin

```
visits          = load   /ipldata/visits   as (user,matchid, time);

gMatches     = group visits by matchId;

matchPopularity  = foreach gMatches generate matchId,
count(visits);


matchInfo        = load   /ipldata/matchInfo   as (url, venue,
winner);

venueCounts  = join gMatches by matchId, matchInfo by
matchId;

gVenues = group venueCounts by venue;

topMatches = foreach gVenues generate
top(matchPopularity,10);


store topMatches into   /data/topMatches  ;
```



# BIG DATA

## Pig References

T1 Chapter 4.6 you are not expected to memorize the syntax of Pig. Just the basics and how it is converted to map reduce tasks. Please go through the entire section if you want to learn to code.

T2 Chapter 7.5

# SQOOP

## Why Sqoop ?

Sometimes we need to use data periodically from a
- Data warehouse
- SQL database

For performing analytics

And store the data back into an SQL database

SQOOP SQL to Hadoop provides this functionality

# BIG DATA

## What is SQOOP?

Bulk Data Transfer Tool  voluminous data

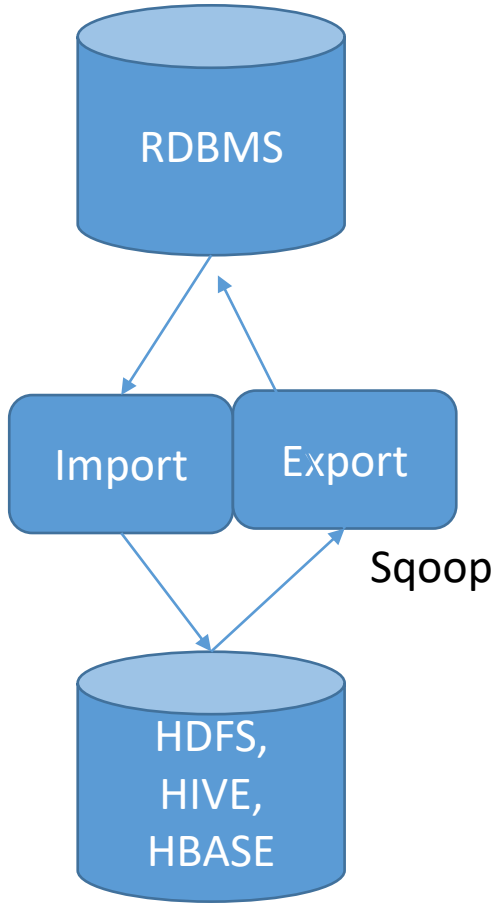
Import/Export data to/from SQL

Defines schema for import

Integrates with Oozie as an action

Support plugins for data sources

supported by default.



# BIG DATA

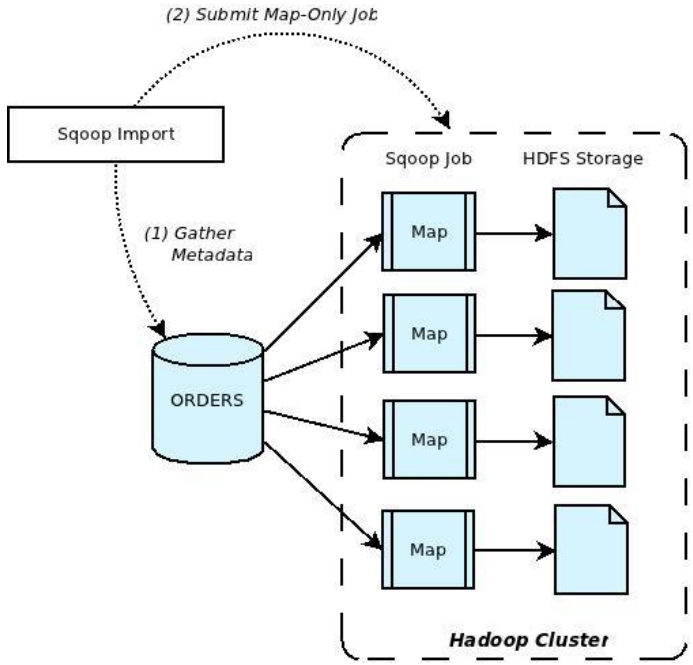
## How does sqoop work? (IMPORT)

Step 1

- Inspects database to gather necessary metadata on data being imported

Step 2

- Transfers the data
  - Map only hadoop job
  - Stores data to hdfs directory
  - Imports into csv file, with newline as record separator

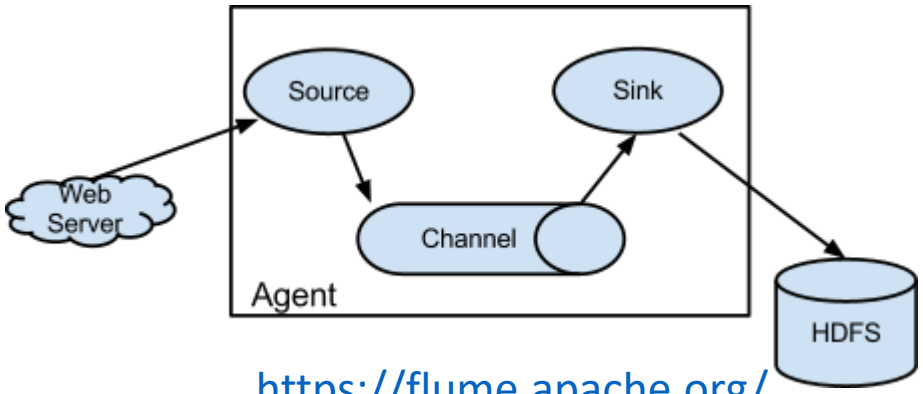



https://blogs.apache.org/sqoop/entry/apache_sqoop_overview

# FLUME

# BIG DATA

## What is Flume ?




https://flume.apache.org/

- Meant for collecting large amounts of streaming data
  - Events
  - Logs like web server logs
- Architecture
  - Sources accept data from an application
  - Sinks receive data and store into HDFS
  - Channels connect sources to sinks
  - Agents run the sources and sinks within Flume

# THANK YOU

**K V Subramaniam**
Dept. of Computer Science and Engineering

**subramaniamkv@pes.edu**