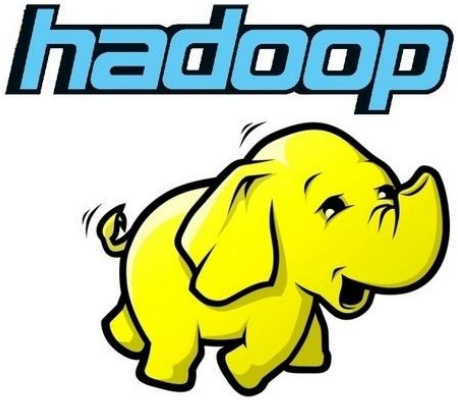
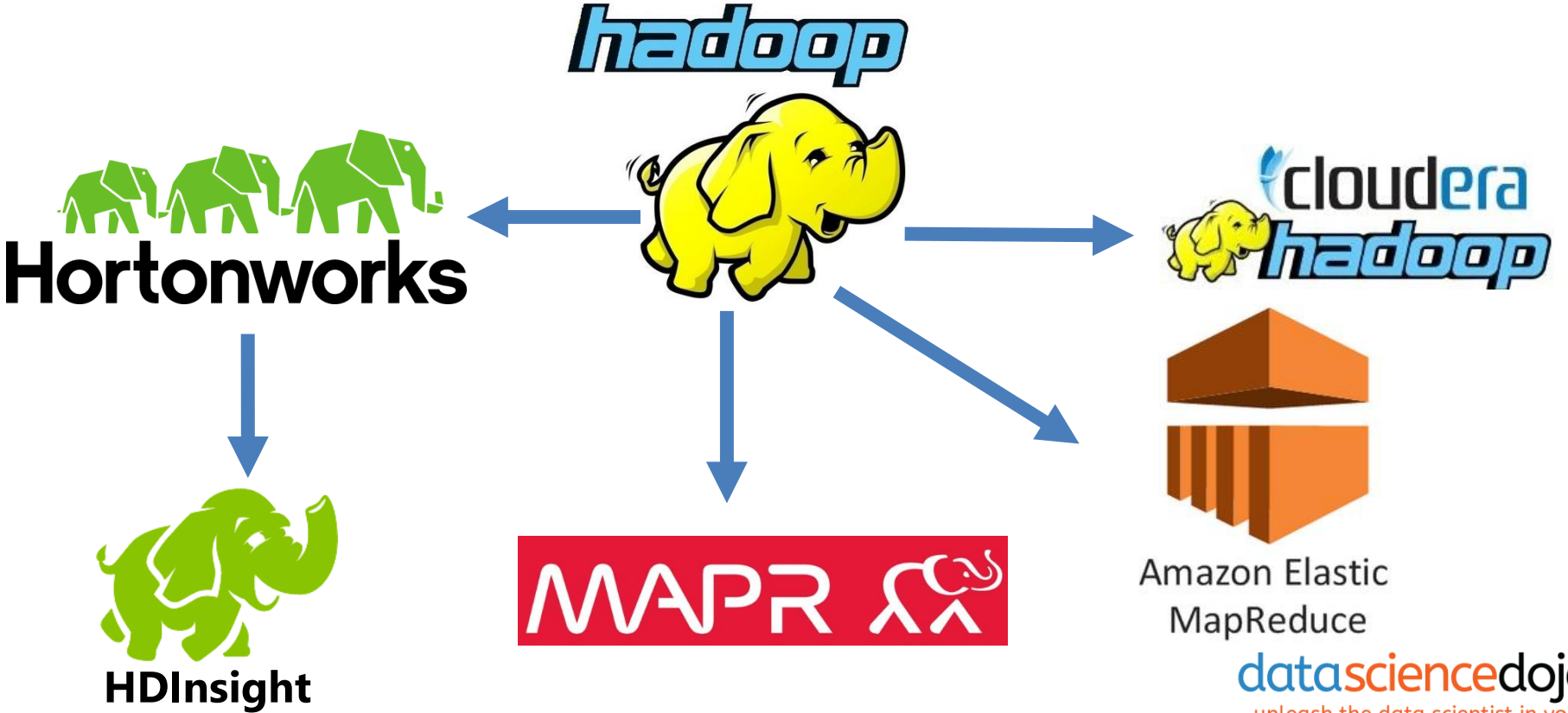
Big Data Engineering With MapReduce and Hive

Data Science Dojo

Agenda

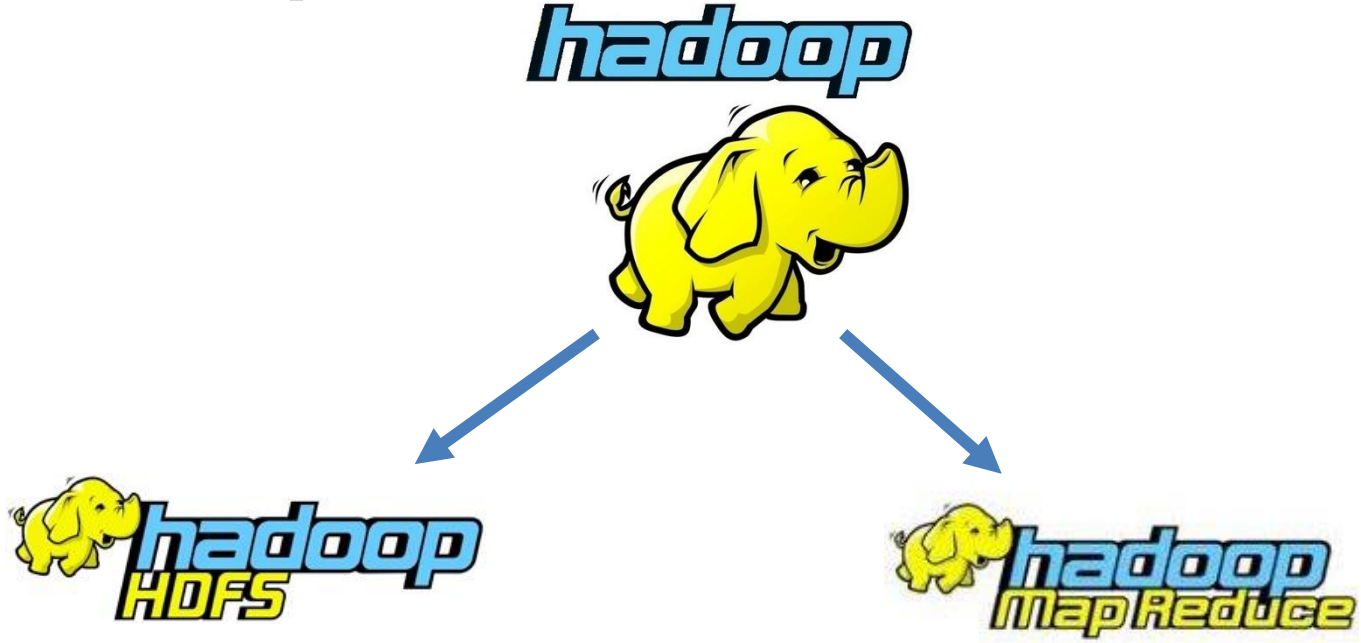


Hadoop Implementations



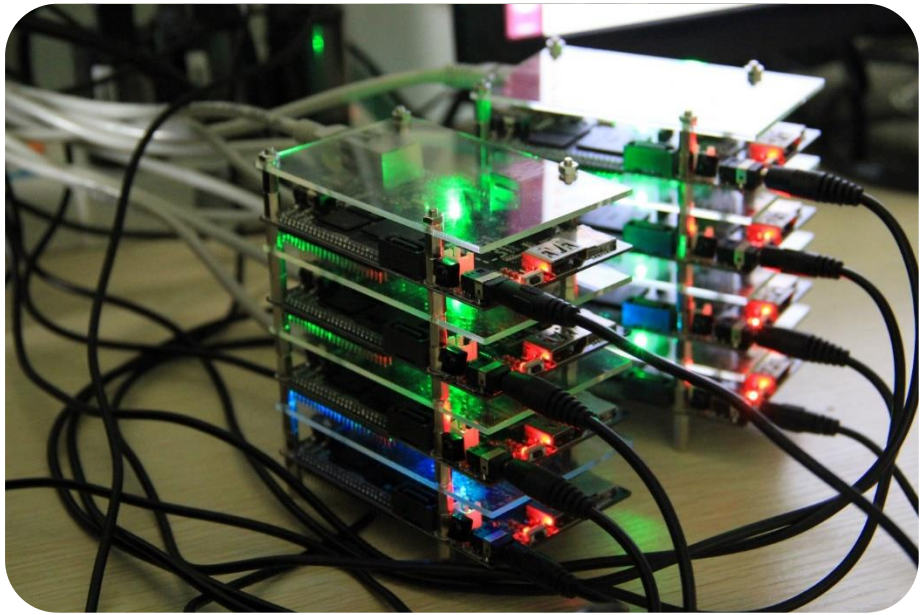
datasciencedojo
unleash the data scientist in you

Hadoop

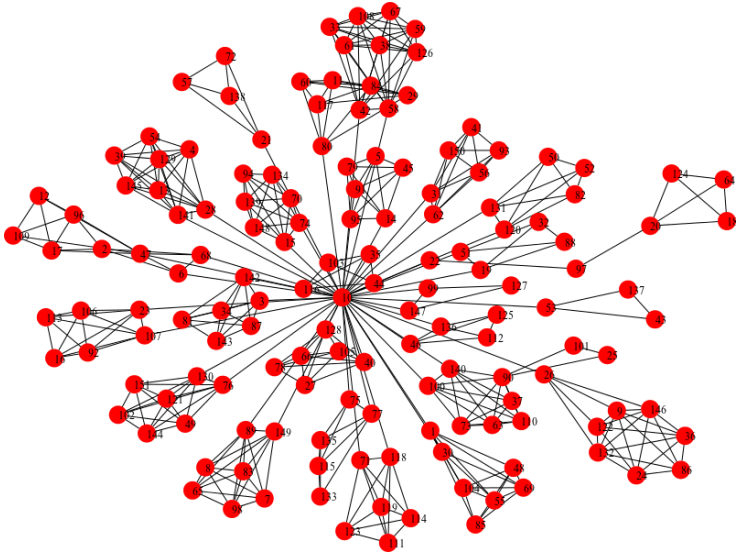


Turn Back The Clock, The Mainframe

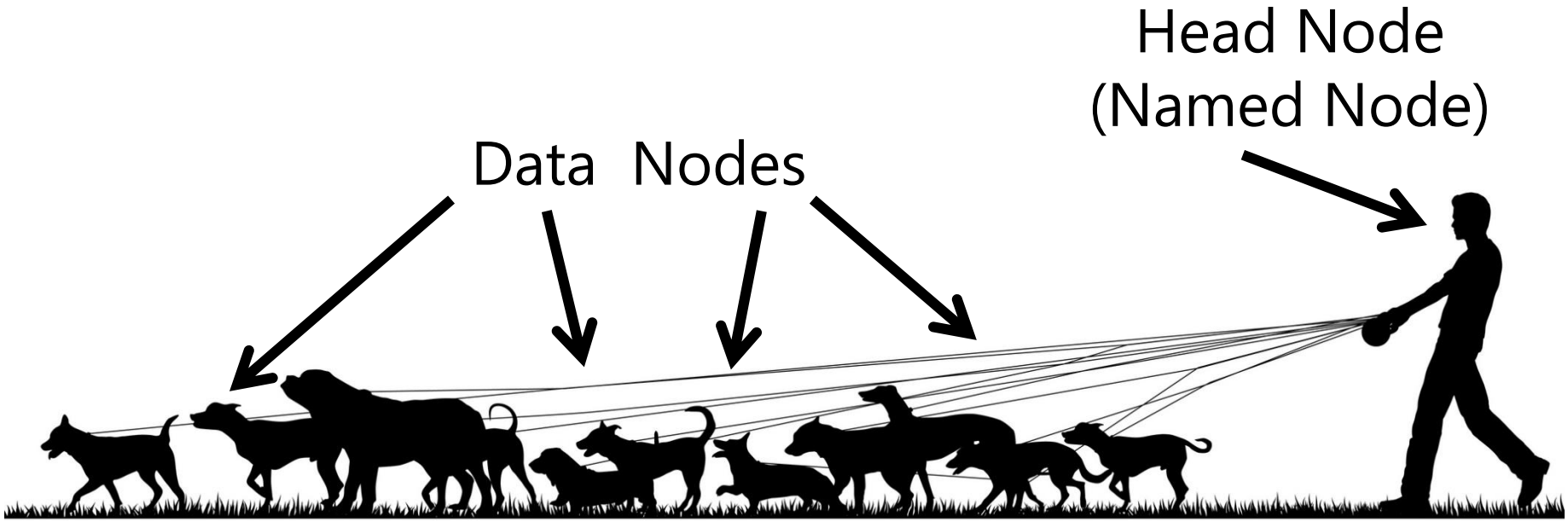
Distributed Computing



Cloud Computing



If dogs were servers…



HDFS & MapReduce

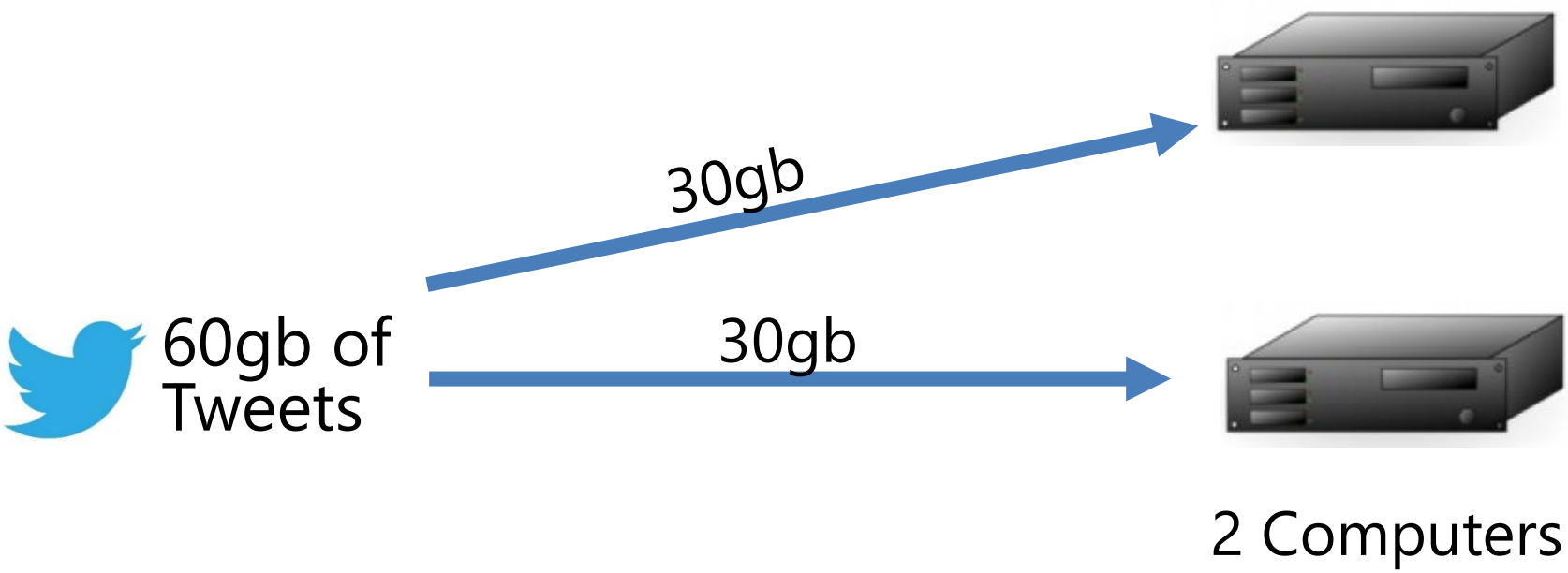
60gb of Tweets

60gb

1 Computer

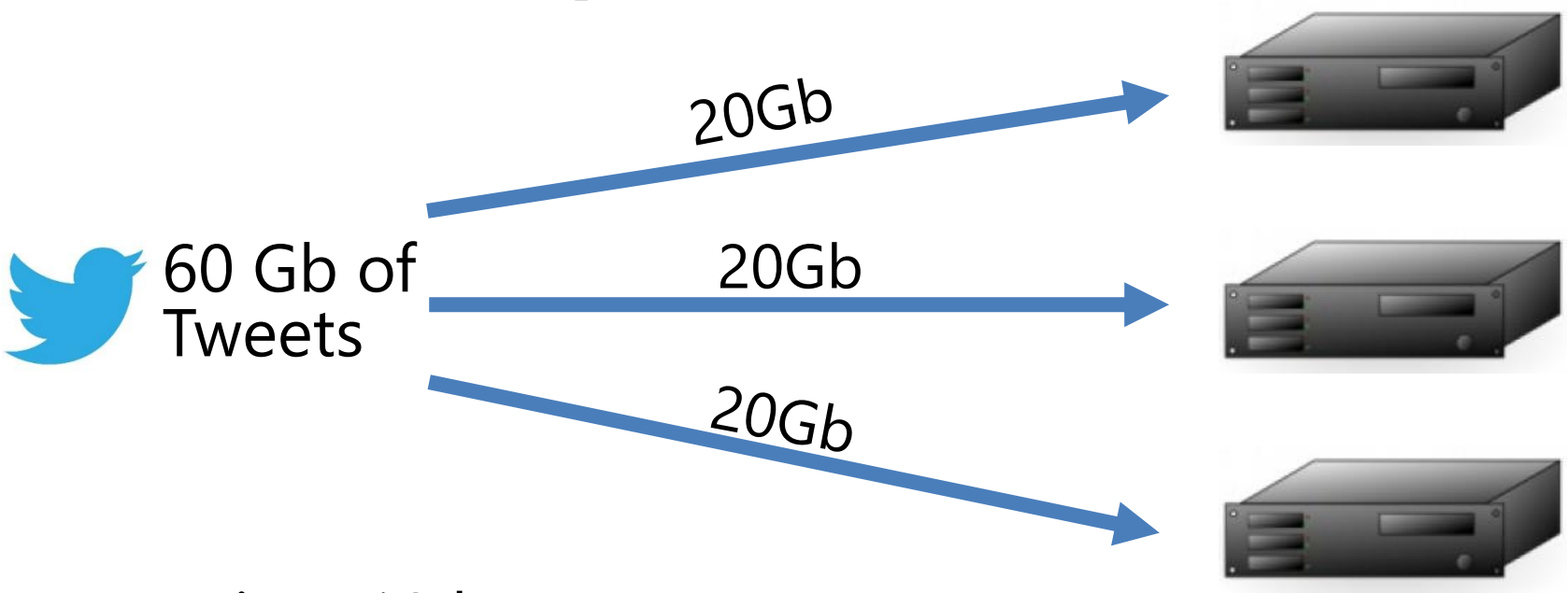
Processing: 30 hours

HDFS & MapReduce



Processing: 15 hours

HDFS & MapReduce



Processing: 10 hours

Most Cases, Linear Scaling Of Processing Power

Number of Computers	Processing Time (hours)
1	30
2	15
3	10
4	7.5
5	6
6	5
7	4.26
8	3.75
9	3.33

Limitations with MapReduce

- ~200 lines of code to do anything
- Slow
- Troubleshooting multiple computers
- Good devs are scarce
- Expensive certifications

```
package org.apache.hadoop.examples;

import java.io.IOException;
import java.util.StringTokenizer;

import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import org.apache.hadoop.util.GenericOptionsParser;

public class WordCount {

  public static class TokenizerMapper
       extends Mapper<Object, Text, Text, IntWritable>{

    private final static IntWritable one = new IntWritable(1);
    private Text word = new Text();

    public void map(Object key, Text value, Context context
                    ) throws IOException, InterruptedException {
      StringTokenizer itr = new StringTokenizer(value.toString());
      while (itr.hasMoreTokens()) {
        word.set(itr.nextToken());
        context.write(word, one);
      }
    }
  }
```

Ambari: Cluster provisioning, management, and monitoring

Avro (Microsoft .NET Library for Avro): Data serialization for the Microsoft .NET environment

HBase: Non-relational database for very large tables

HDFS: Hadoop Distributed File System

Hive: SQL-like querying

Mahout: Machine learning

MapReduce and YARN: Distributed processing and resource management

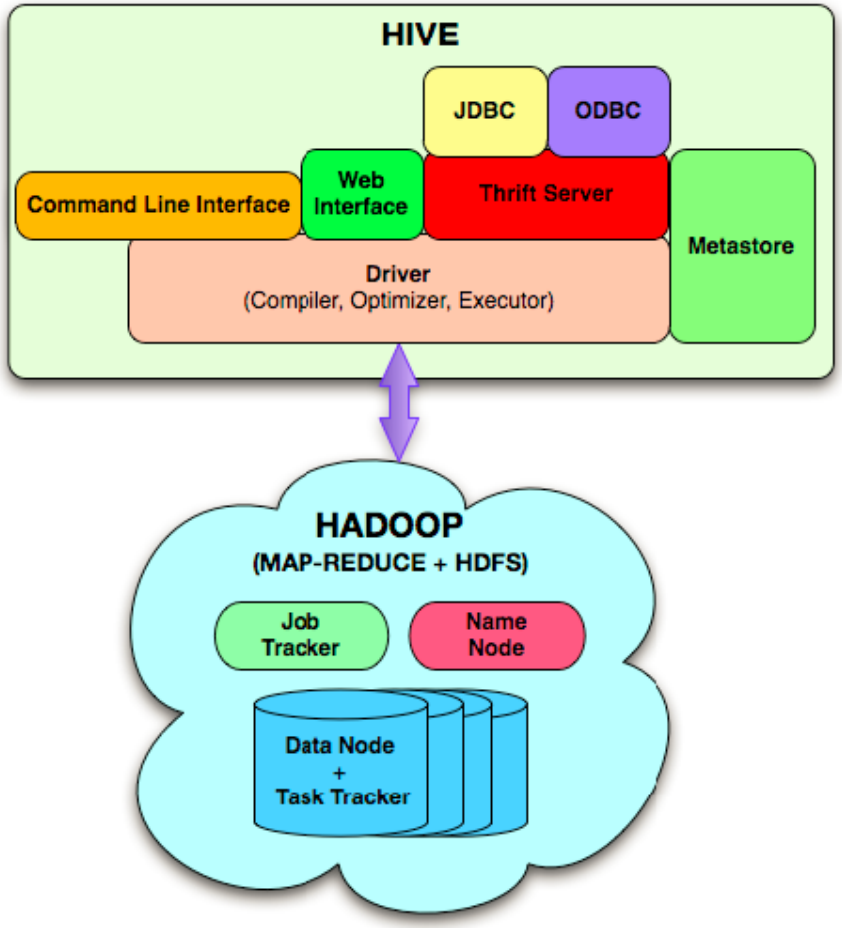
Oozie: Workflow management

Pig: Simpler scripting for MapReduce transformations

Sqoop: Data import and export

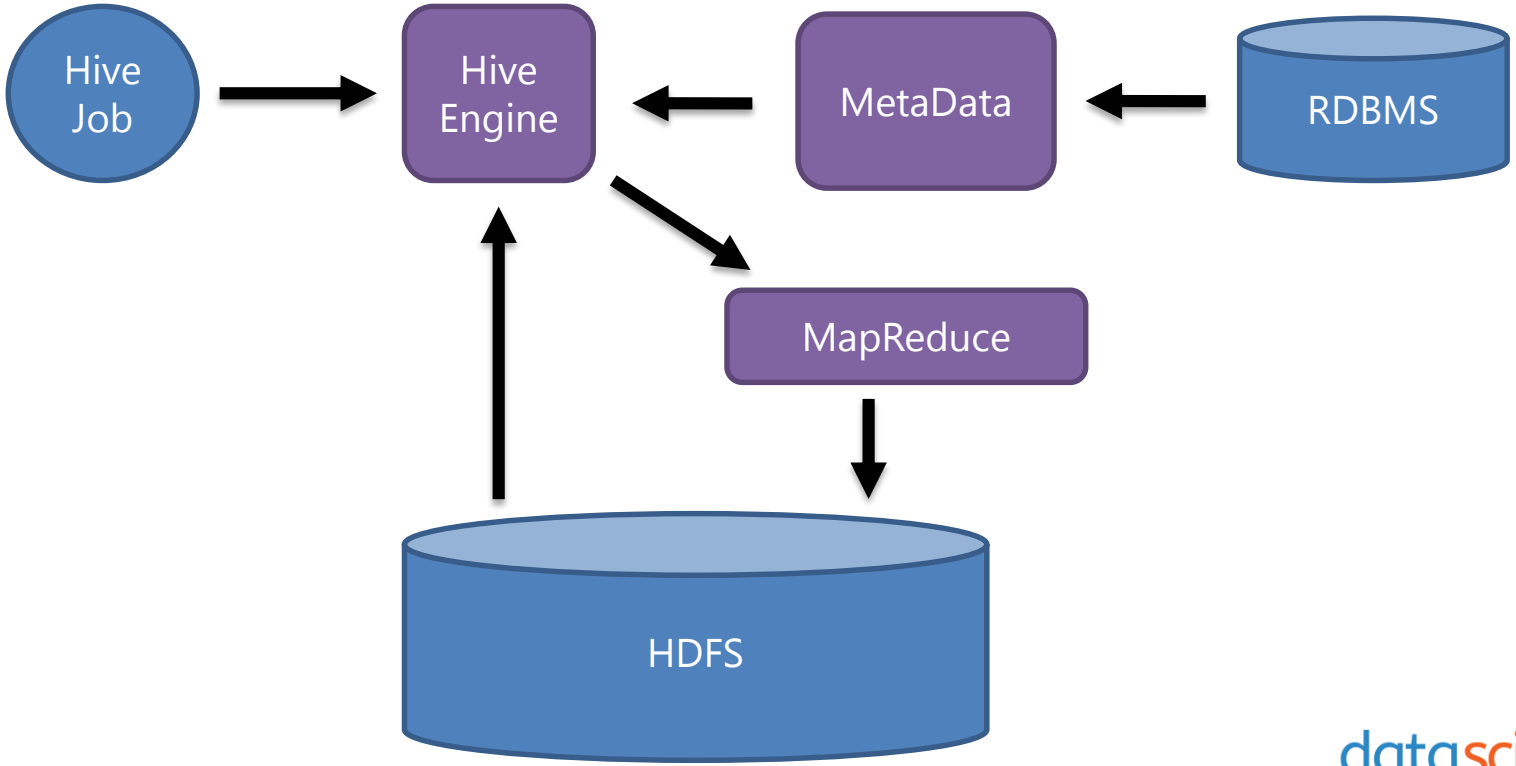
Storm: Real-time processing of fast, large data streams

Zookeeper: Coordinates processes in distributed systems



Hive Jobs

Hive Architecture



Data File

=

Unstructured Data

Data File

+

Metadata File/DB

=

Structured Data

Semi Structured Data

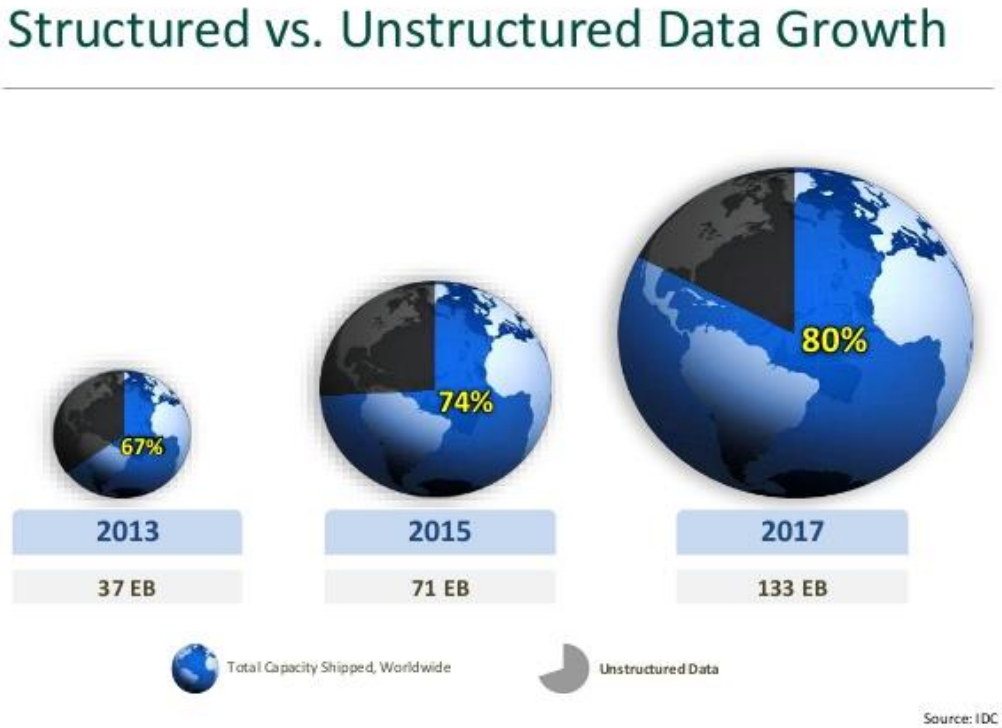
Self Describing Flat Files

- XML
- JSON
- CSV
- TSV

Why Hive?

- SQL spoken here (HiveQL)
- ODBC driver
- BI Integration
- Supports only Structured Data

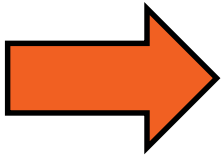
Limitations

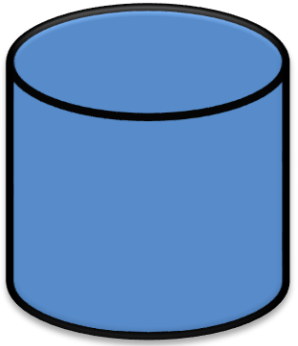


datasciencedojo
unleash the data scientist in you

Azure Blob Storage

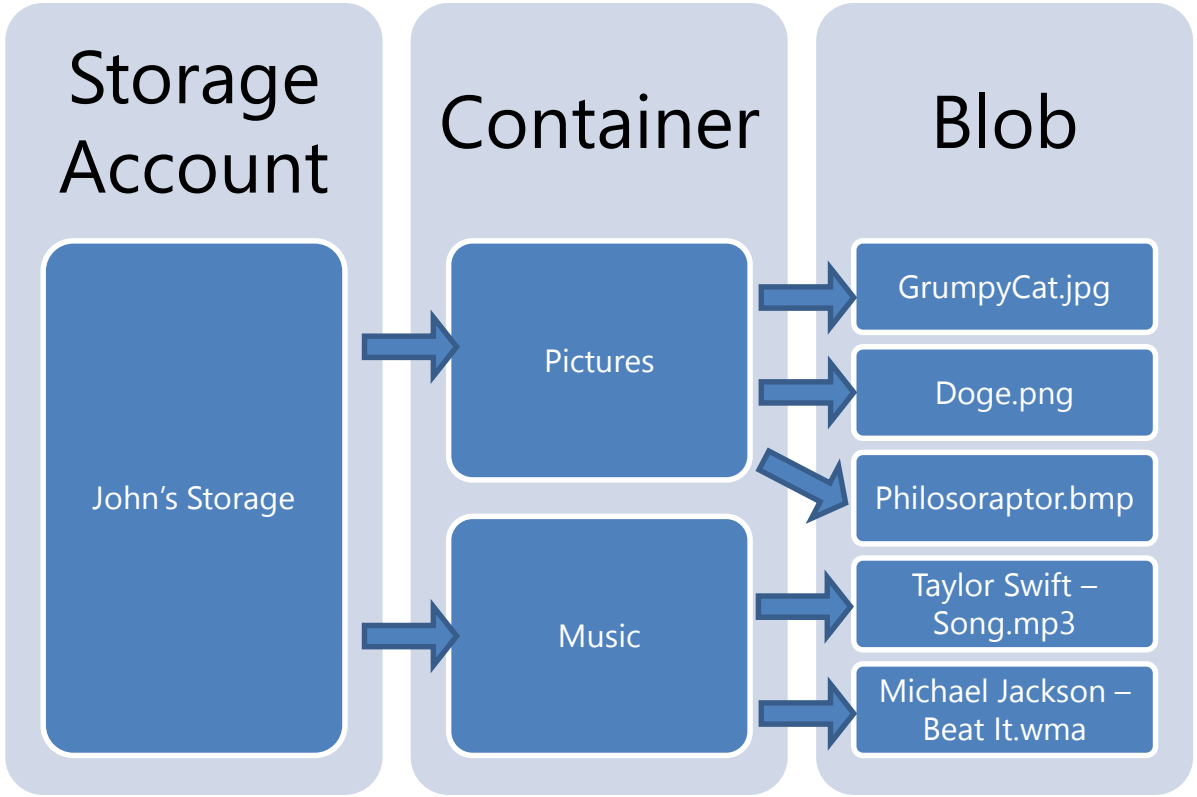
HDInsight



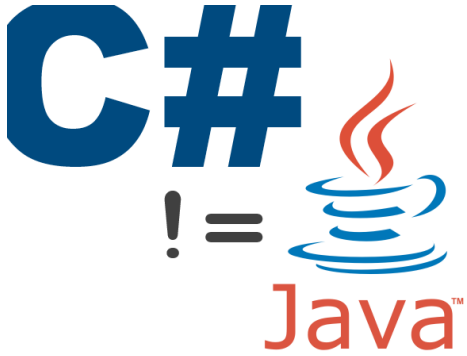


Blob Storage

Azure Blob Storage

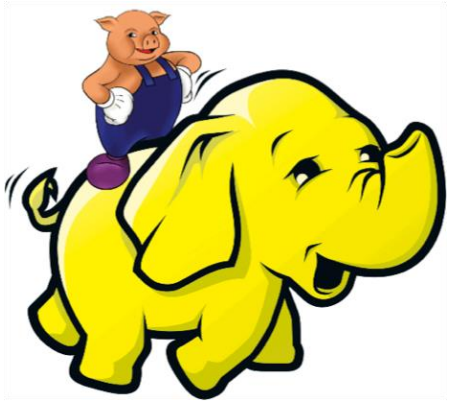


When to Use Each



vs

vs



C#
Java
MapReduce

Hive

Pig

QUESTIONS