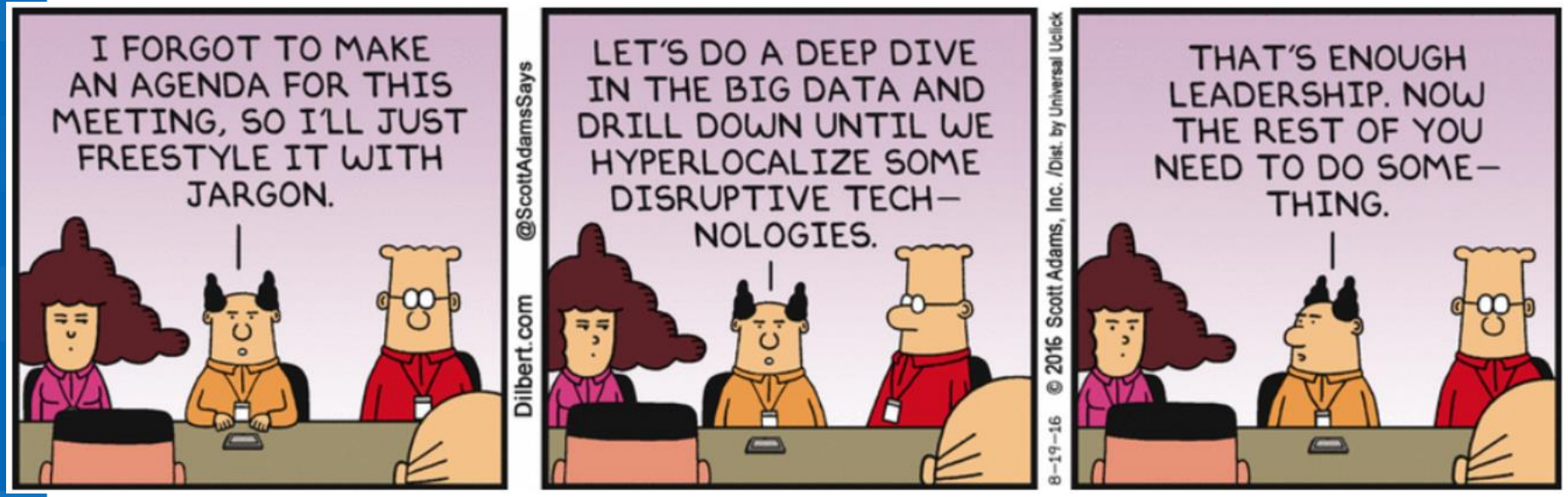
EMR 5.3.1
(Amazon Elastic MapReduce)

Version : 1 [Created By: Pawan S. Mude]

Sterling Talent Solutions

What's Agenda?



Sterling
Talent Solutions

Agenda

- Use Cases
- What is EMR?
- Data Processing Engine behind EMR
- Hadoop Framework
- MapReduce Job Tasks
- EMR Nodes
- Storage Options
- How can I access EMR?
- Get your hands Dirty!
- Can I breath without smelling JAVA?
- Programming Languages Supported
- Where Do I look forward?
- Recap
- What's coming in upcoming trainings?

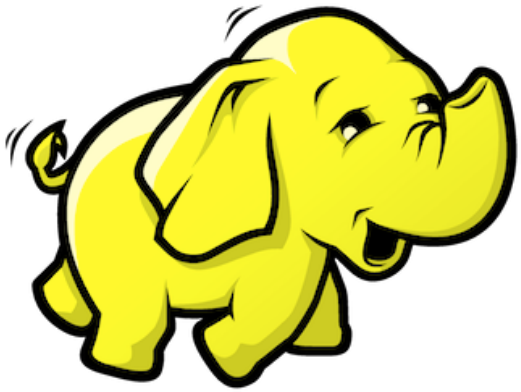
Use Cases ?

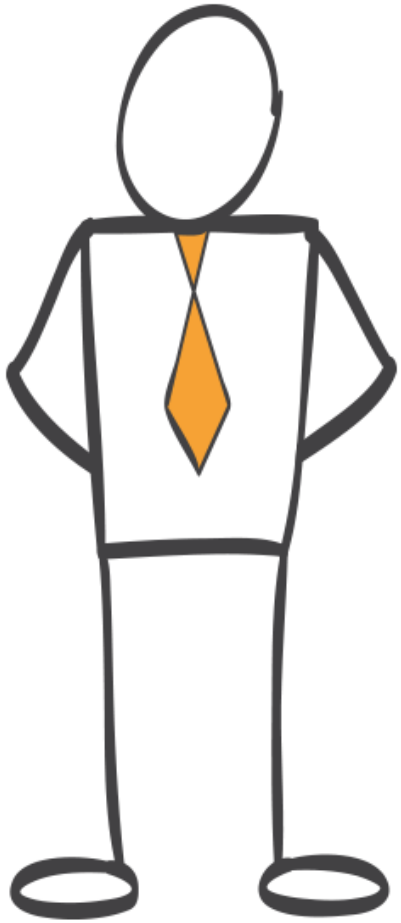
- Real time insights
- Log Processing
- ETL
- ClickStream Analysis
- Machine Learning

What is EMR?

- Managed Cluster Platform.
- WebService that enables businesses, researchers, data analyst and developers to easily and cost-effectively **process** vast amount of data.
- It utilizes Hadoop framework running on web-scale infrastructure of Amazon Elastic Compute Cloud (EC2) and Amazon Simple Storage Service (S3).
- Instantly provision capacity.
- Set-up; Management or tuning of Hadoop cluster & Compute capacity upon which they sit.
- Unique Identifier start with “j-”.

Data Processing Engine behind EMR





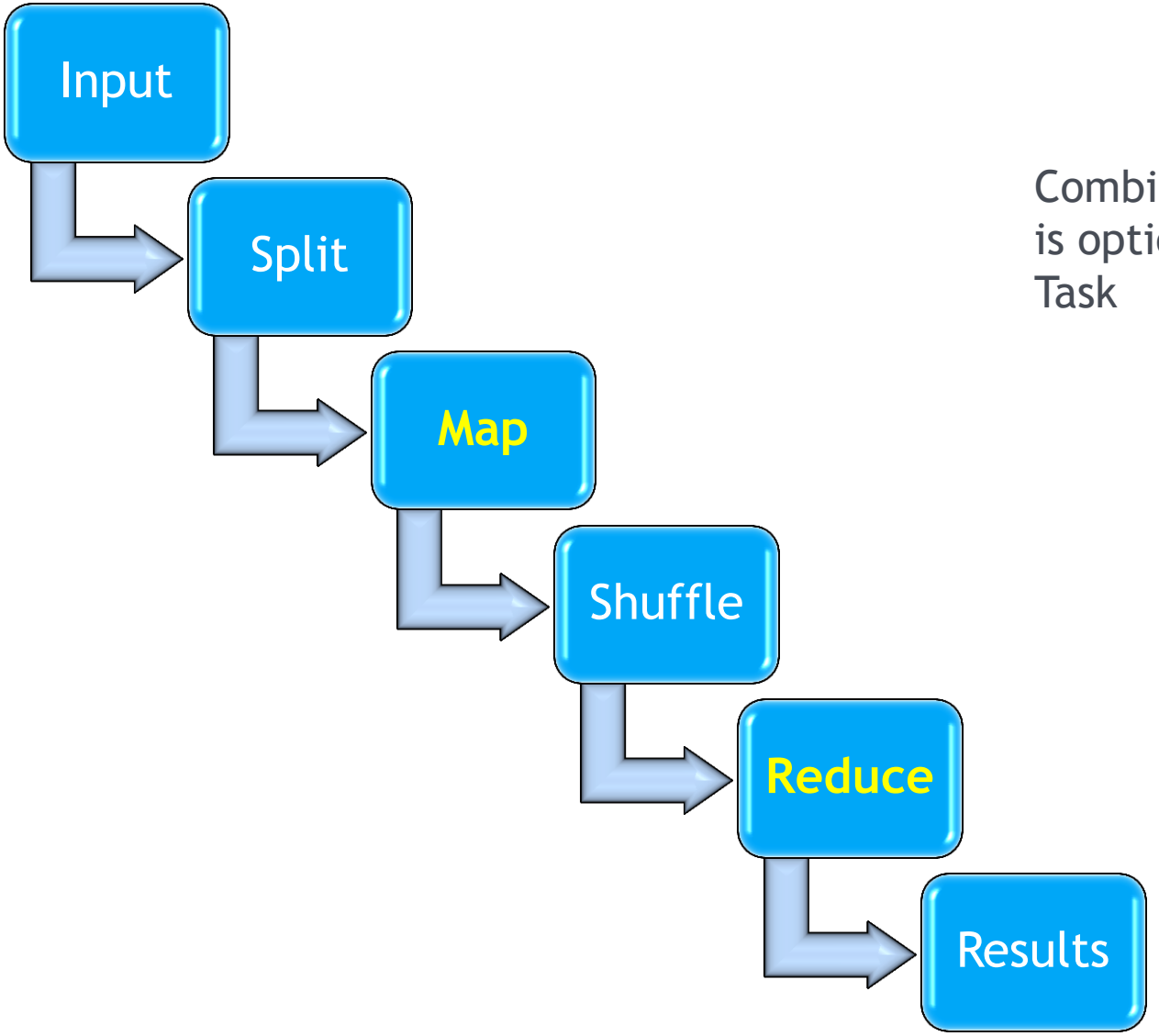
Hadoop Framework

- Hadoop Common (core)
- HDFS
- YARN
- MapReduce (Parallel, distributed)

Same old Political Strategy: Divide & Conquer

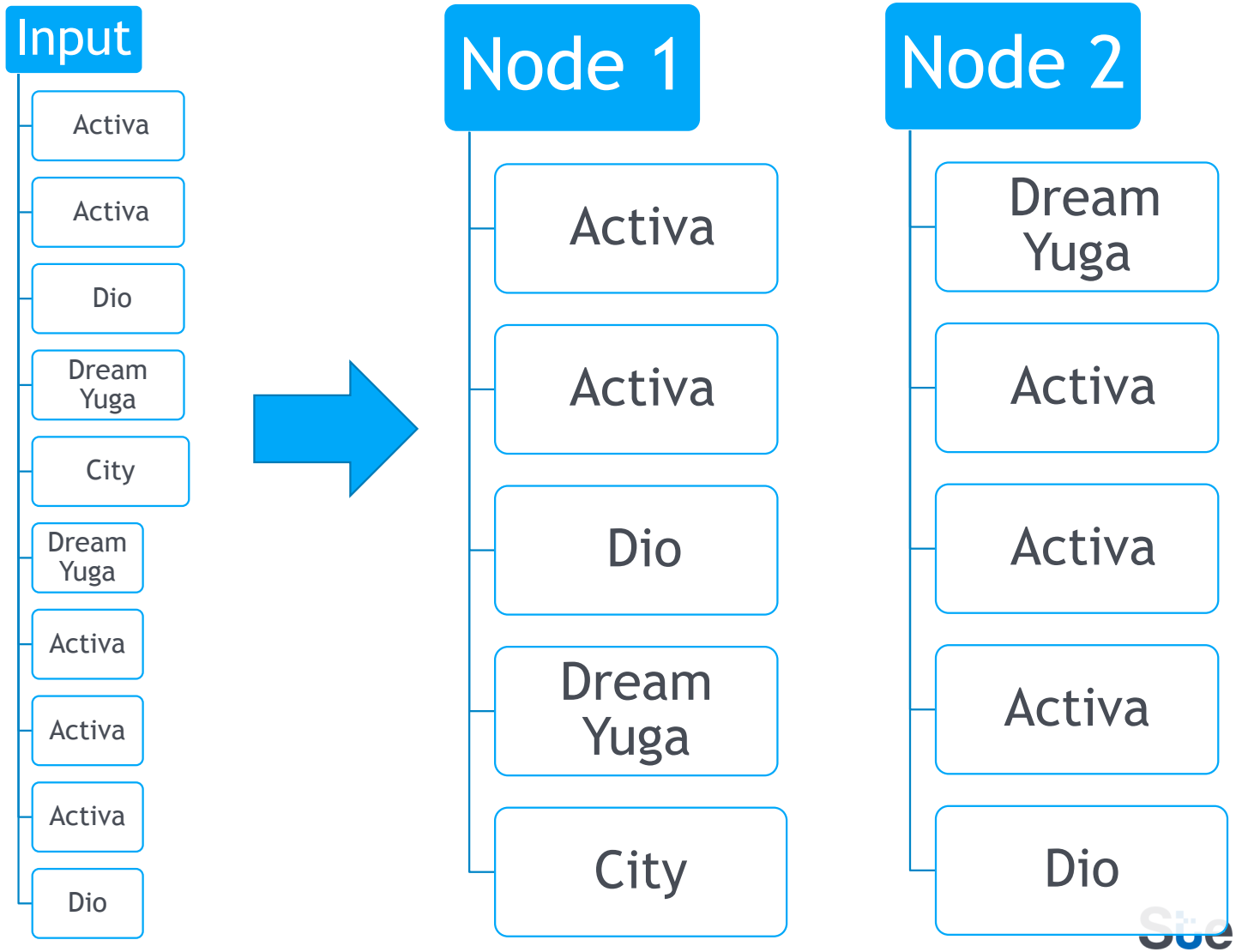
Sterling
Talent Solutions

MapReduce Job Tasks

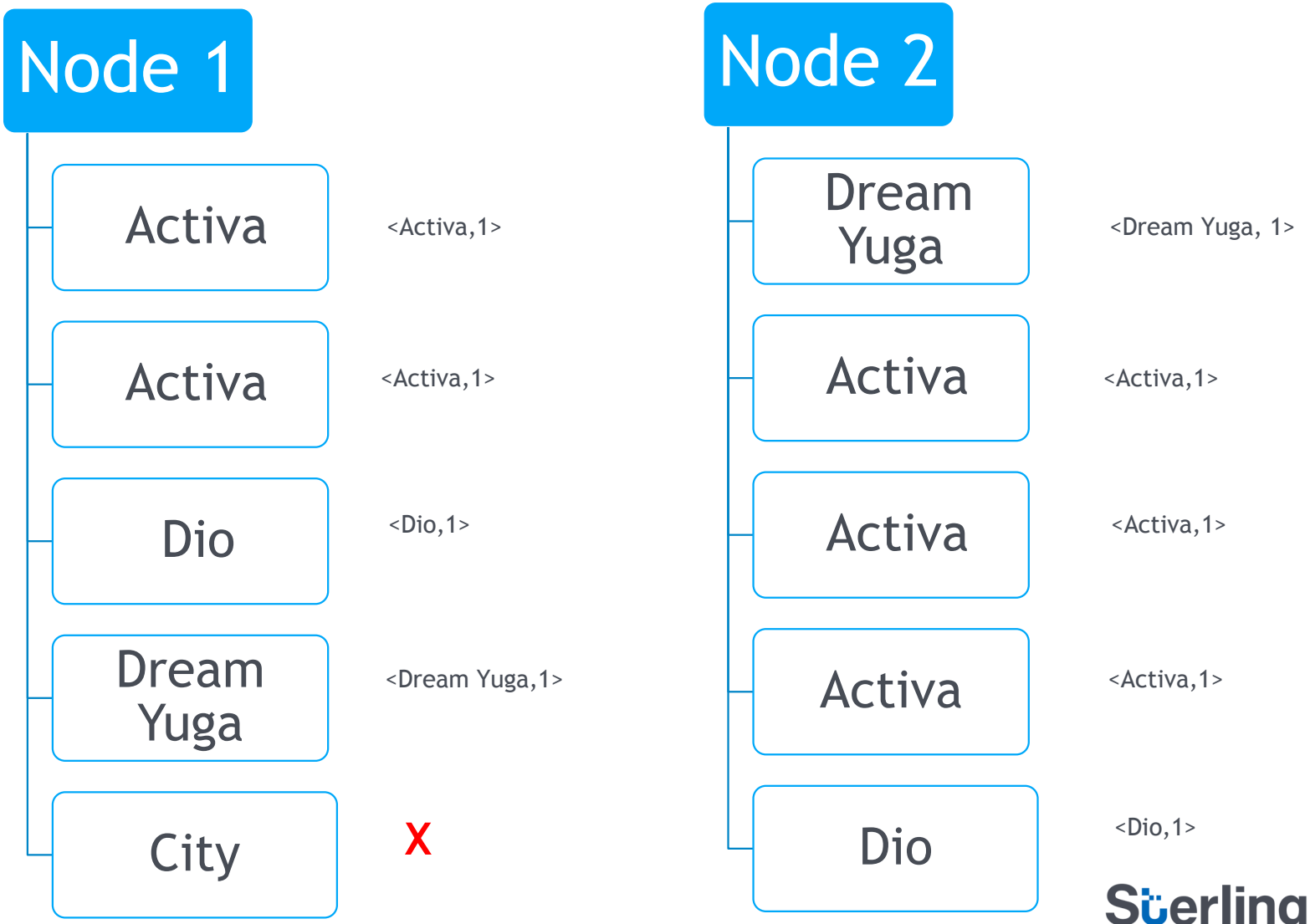


Combine
is optional
Task

About Map Reduce : Split

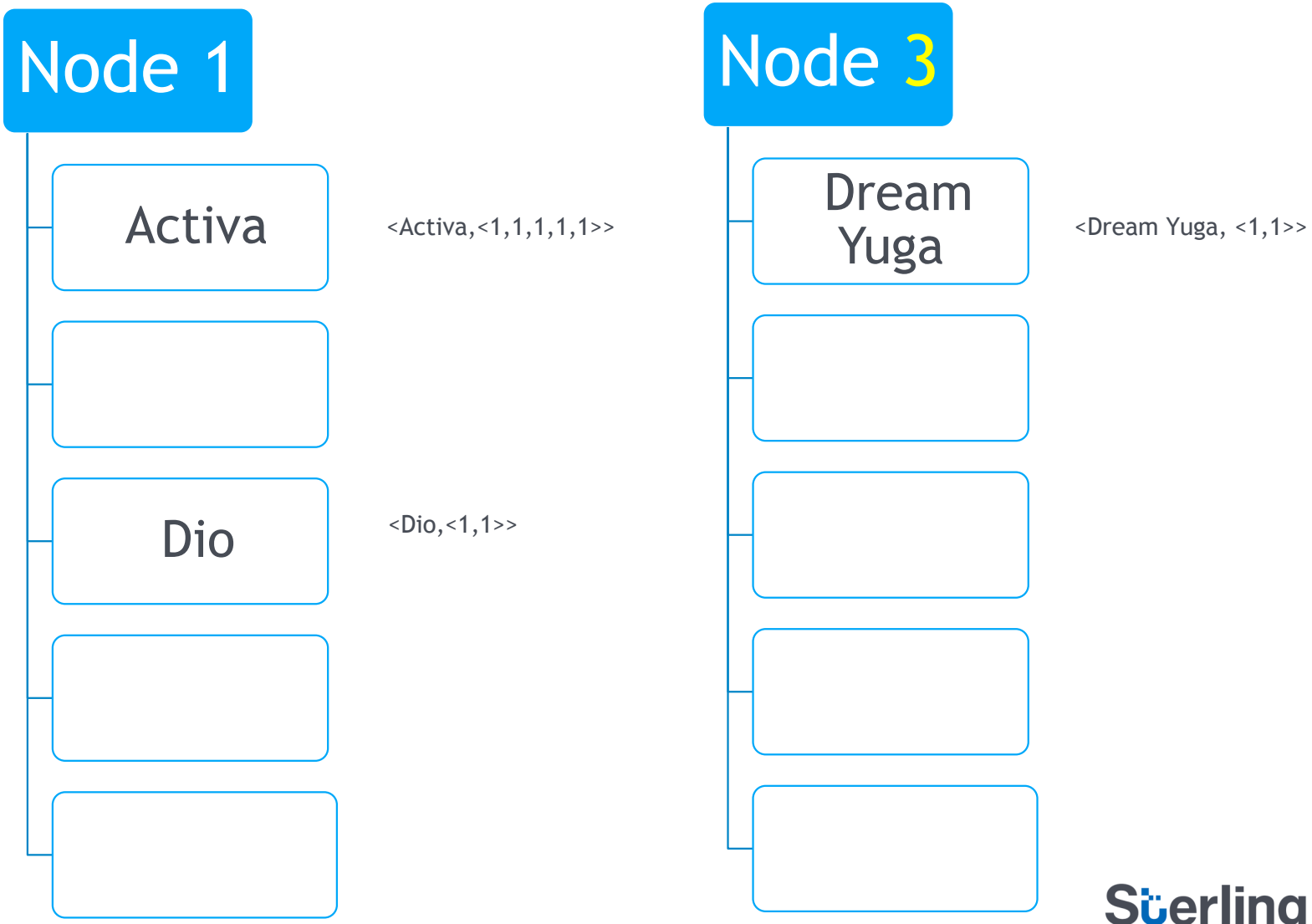


About Map Reduce : Map

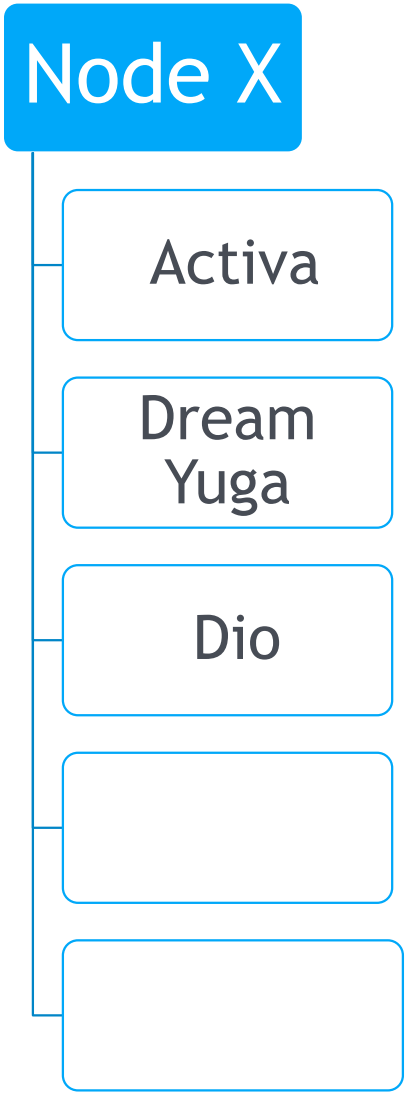


Sterling Talent Solutions

About Map Reduce : Shuffle



About Map Reduce : Reducer Result



<Activa,<5>>

<Dream Yuga, <2>>

<Dio,<2>>

Note:
Combine is optional Phase.

Sterling Talent Solutions

EMR Nodes

- Master Node : Manages Resources of Cluster (Single, Coordinates Distribution & Parallel execution. Metadada).
- Core Nodes : Run Tasks. Store Data. DataNode daemons runs of core nodes.
- Task Nodes: Optional . Can be added or removed to for extra CPU/RAM.

Storage Options in EMR

- Instance Store
- EBS
- EMR File System (EMRFS)

How can I access EMR?

- AWS Management console.
- Command Line Tools.
- SDKs.
- EMR API.

Get your hands Dirty!

```
[hadoop@ip-10-60-7-65 labfiles]$  hadoop jar /home/hadoop/testfile/MaxMT.jar MaxMonthTemp sampledata/SumnerCountyTemp.dat sampledata/TempOut2
17/07/07 04:41:27 INFO impl.TimelineClientImpl: Timeline service address: http://ip-10-60-7-65.ec2.internal:8188/ws/v1/timeline/
17/07/07 04:41:27 INFO client.RMProxy: Connecting to ResourceManager at ip-10-60-7-65.ec2.internal/10.60.7.65:8032
17/07/07 04:41:27 INFO input.FileInputFormat: Total input paths to process : 1
17/07/07 04:41:27 INFO lzo.GPLNativeCodeLoader: Loaded native gpl library
17/07/07 04:41:27 INFO lzo.LzoCodec: Successfully loaded & initialized native-lzo library [hadoop-lzo rev 60b8618a21bca805756fb1bc393c18c2512e4fc4]
17/07/07 04:41:28 INFO mapreduce.JobSubmitter: number of splits:1
17/07/07 04:41:28 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1486999765918_0895
17/07/07 04:41:28 INFO impl.YarnClientImpl: Submitted application application_1486999765918_0895
17/07/07 04:41:28 INFO mapreduce.Job: The url to track the job: http://ip-10-60-7-65.ec2.internal:20888/proxy/application_1486999765918_0895/
17/07/07 04:41:28 INFO mapreduce.Job: Running job: job_1486999765918_0895
17/07/07 04:41:35 INFO mapreduce.Job: Job job_1486999765918_0895 running in uber mode : false
17/07/07 04:41:35 INFO mapreduce.Job:  map 0% reduce 0%
17/07/07 04:41:40 INFO mapreduce.Job:  map 100% reduce 0%
17/07/07 04:41:46 INFO mapreduce.Job:  map 100% reduce 13%
17/07/07 04:41:49 INFO mapreduce.Job:  map 100% reduce 20%
17/07/07 04:41:50 INFO mapreduce.Job:  map 100% reduce 27%
17/07/07 04:41:52 INFO mapreduce.Job:  map 100% reduce 33%
17/07/07 04:41:53 INFO mapreduce.Job:  map 100% reduce 40%
17/07/07 04:41:55 INFO mapreduce.Job:  map 100% reduce 47%
17/07/07 04:41:56 INFO mapreduce.Job:  map 100% reduce 53%
17/07/07 04:41:58 INFO mapreduce.Job:  map 100% reduce 60%
17/07/07 04:41:59 INFO mapreduce.Job:  map 100% reduce 67%
17/07/07 04:42:01 INFO mapreduce.Job:  map 100% reduce 73%
17/07/07 04:42:03 INFO mapreduce.Job:  map 100% reduce 80%
17/07/07 04:42:04 INFO mapreduce.Job:  map 100% reduce 87%
17/07/07 04:42:07 INFO mapreduce.Job:  map 100% reduce 93%
17/07/07 04:42:08 INFO mapreduce.Job:  map 100% reduce 100%
17/07/07 04:42:08 INFO mapreduce.Job: Job job_1486999765918_0895 completed successfully
17/07/07 04:42:08 INFO mapreduce.Job: Counters: 50
```

Sterling
Talent Solutions

Can I breath without smelling JAVA?

- Hadoop is implemented in JAVA.
- Hive is useful ‘High-Level’ programming language, for non-java programmers. [Apache Tez is default execution engine for hive instead of mapreduce].
- Hive provides SQL abstraction to integrate HiveQL queries into the underlying JAVA API without having to write JAVA programs.
- Presto (in-memory fast SQL Query Engine); does not need interpretor like Hive. (Not suitable for Batch, 100M+ rows).

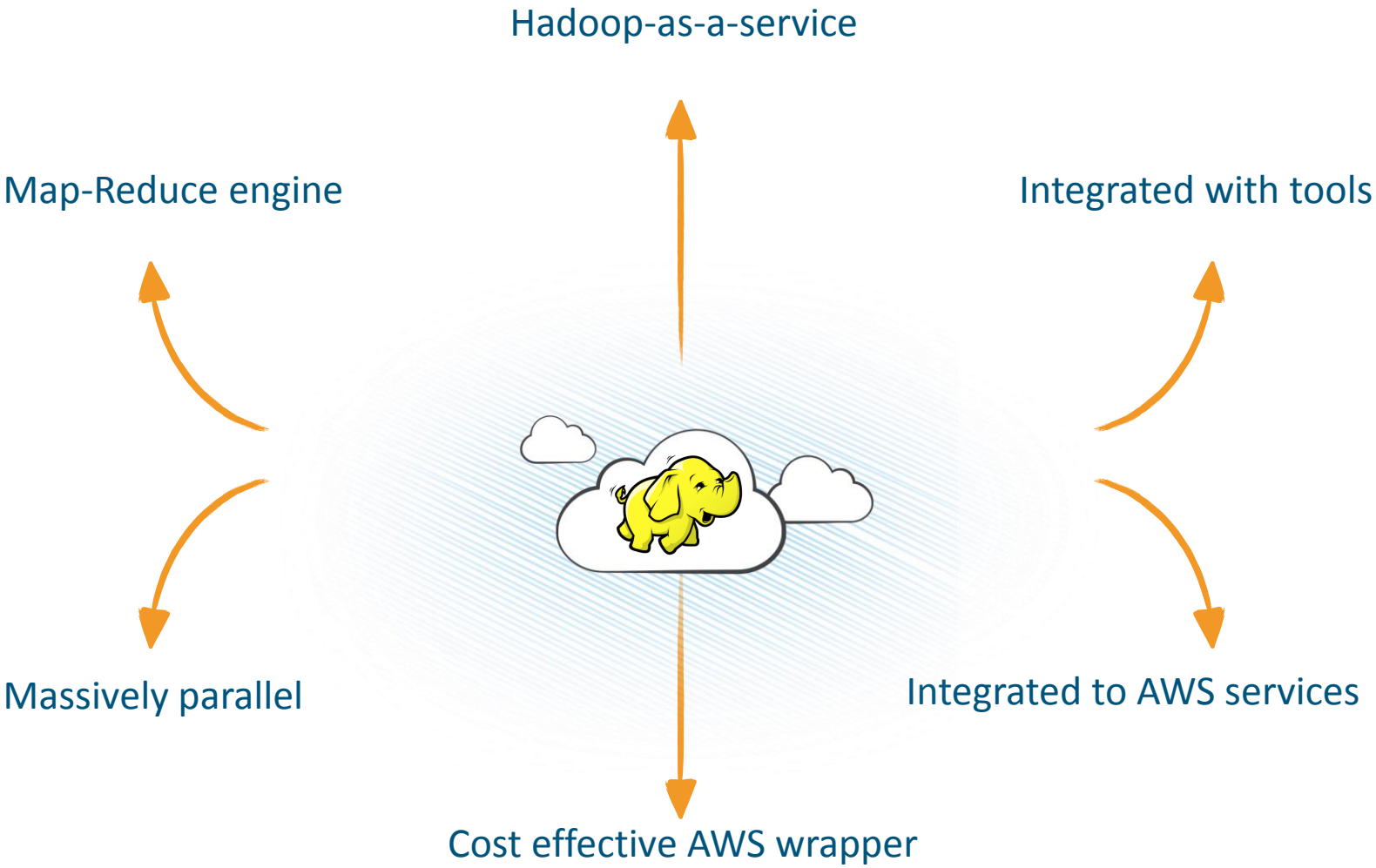
Programming Languages Supported

- Programming Languages Supported :
  - Java (to implement custom JARs)
  - Perl
  - Python
  - Ruby
  - C++
  - PHP
  - R via Hadoop Streaming

Where Do I look forward?

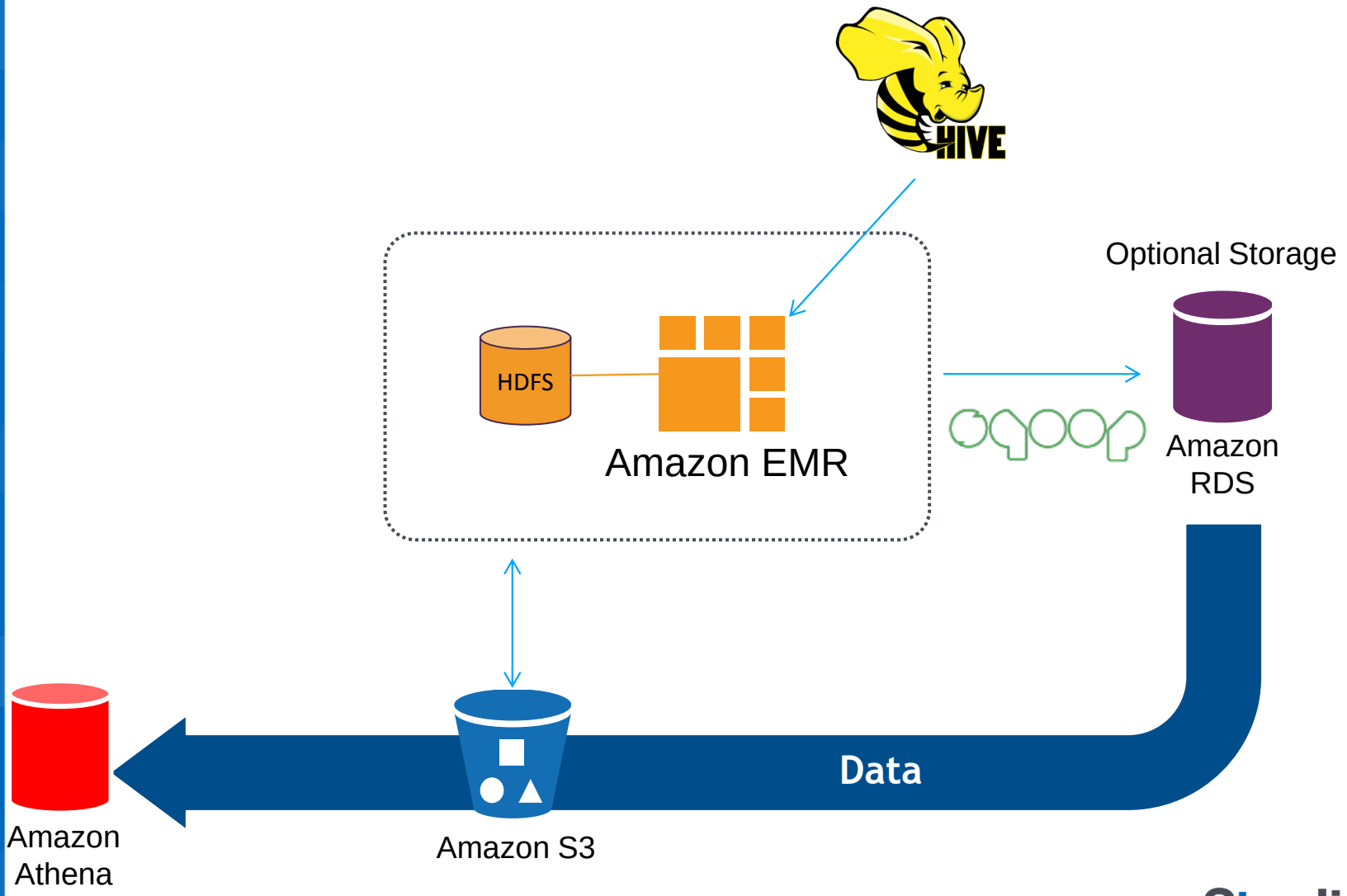
- Apache Spark
  - Fast Engine for processing large amount of Data.
  - Run in-memory
  - Run on disk
- Spark for Interactive Analytics; faster than MapReduce.
- Run Quires against live data.
- Flexibility in terms of languages used (Scala, Python, etc).
- Good Solution to train Machine Learning algorithms.
- Not good for Batch Processing.

Recap



Sterling
Talent Solutions

What's coming in upcoming trainings?



Sterling
Talent Solutions

Get your hands dirty with hand-gloves!

2. 10.60.7.65 (hadoop)

```
[hadoop@ip-10-60-7-65 ~]$ sqoop import --connect jdbc:mysql://smartdata-dev.ctwaaurutcz2.us-east-1.rds.amazonaws.com/smartdata_dev --username smartdata -P --table test --as-parquetfile --target-dir /tmp/pawan --split-by jurisdiction
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
17/07/07 08:56:33 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6
Enter password:
17/07/07 08:56:38 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
17/07/07 08:56:38 INFO tool.CodeGenTool: Beginning code generation
17/07/07 08:56:38 INFO tool.CodeGenTool: Will generate java class as codegen_test
17/07/07 08:56:38 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `test` AS t LIMIT 1
17/07/07 08:56:38 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `test` AS t LIMIT 1
17/07/07 08:56:38 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-hadoop/compile/c1774567bafb418a641eb8537f3382cd/codegen_test.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
17/07/07 08:56:41 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-hadoop/compile/c1774567bafb418a641eb8537f3382cd/codegen_test.jar
17/07/07 08:56:41 WARN manager.MySQLManager: It looks like you are importing from mysql.
17/07/07 08:56:41 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
17/07/07 08:56:41 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
17/07/07 08:56:41 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
17/07/07 08:56:41 INFO mapreduce.ImportJobBase: Beginning import of test
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/hive/lib/log4j-slf4j-impl-2.4.1.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
17/07/07 08:56:41 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
17/07/07 08:56:42 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `test` AS t LIMIT 1
17/07/07 08:56:42 INFO conf.HiveConf: Found configuration file file:/etc/hive/conf.dist/hive-site.xml
17/07/07 08:56:43 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
17/07/07 08:56:44 INFO impl.TimelineClientImpl: Timeline service address: http://ip-10-60-7-65.ec2.internal:8188/ws/v1/timeline/
17/07/07 08:56:44 INFO client.RMProxy: Connecting to ResourceManager at ip-10-60-7-65.ec2.internal/10.60.7.65:8032
17/07/07 08:56:45 INFO db.DBInputFormat: Using read commited transaction isolation
17/07/07 08:56:45 INFO db.DataDrivenDBInputFormat: BoundingValsQuery: SELECT MIN(`jurisdiction`), MAX(`jurisdiction`) FROM `test`
17/07/07 08:56:45 INFO mapreduce.JobSubmitter: number of splits:4
17/07/07 08:56:45 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1486999765918_0897
17/07/07 08:56:46 INFO impl.YarnClientImpl: Submitted application application_1486999765918_0897
17/07/07 08:56:46 INFO mapreduce.Job: The url to track the job: http://ip-10-60-7-65.ec2.internal:20888/proxy/application_1486999765918_0897/
17/07/07 08:56:46 INFO mapreduce.Job: Running job: job_1486999765918_0897
17/07/07 08:56:54 INFO mapreduce.Job: Job job_1486999765918_0897 running in uber mode : false
17/07/07 08:56:54 INFO mapreduce.Job:  map 0% reduce 0%
17/07/07 08:57:03 INFO mapreduce.Job:  map 75% reduce 0%
17/07/07 08:57:04 INFO mapreduce.Job:  map 100% reduce 0%
```

Sterling Talent Solutions

Thank You for your time !

Sterling
Talent Solutions