

MAPR[®]

Introduction to Apache Hadoop Ecosystem



© 2014 MapR Technologies

This lesson is a short introduction to the Apache Hadoop ecosystem.



Learning Objectives

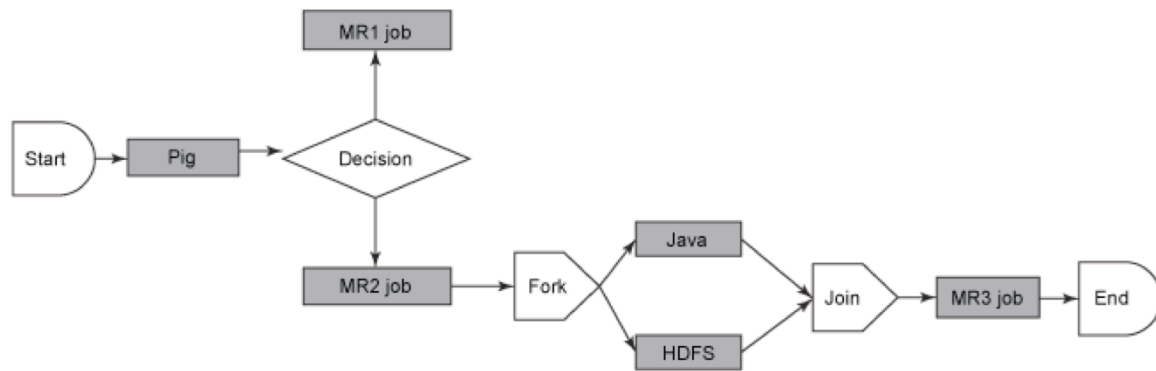
- Discuss the motivation for Oozie
- Describe how to use Oozie
- Perform a basic demonstration of Oozie



The objectives for this lesson are defined above.

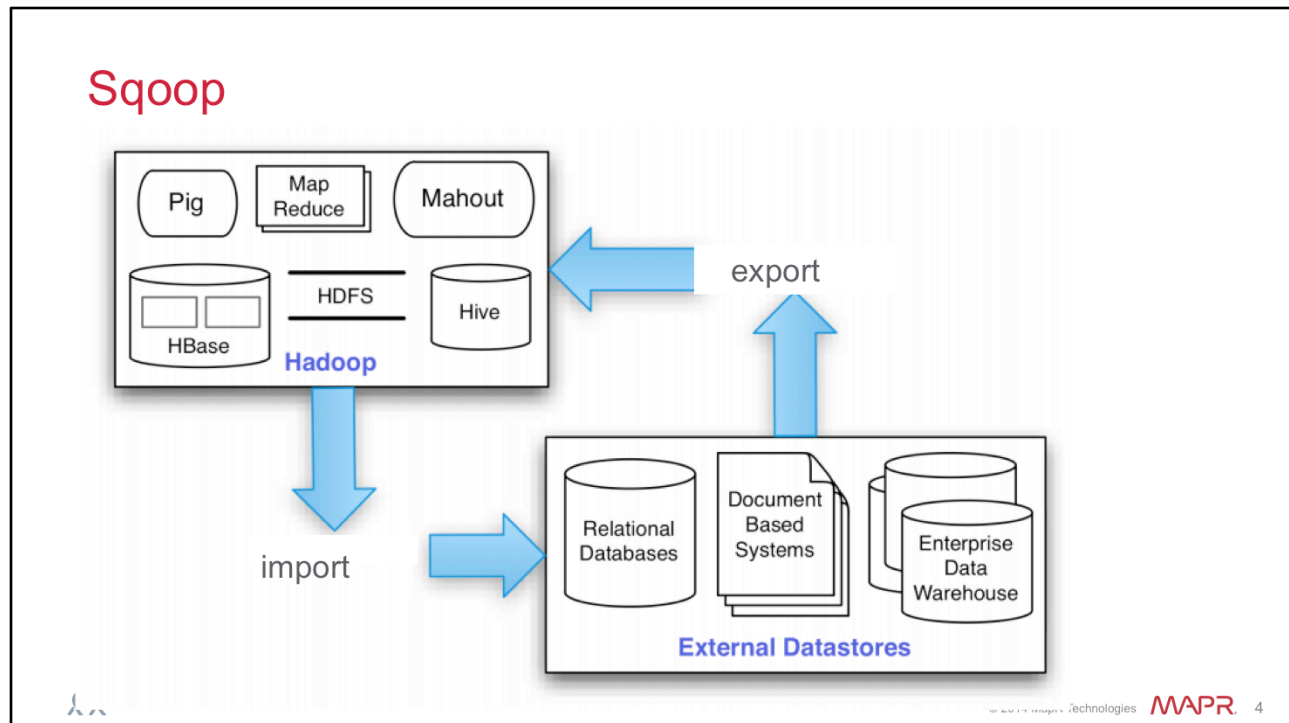


Oozie



The diagram above (called an acyclic directed graph, or DAG) represents a simple Hadoop workflow. How would you manage such a workflow?





A typical data flow in a Hadoop cluster is shown in the slide above. Data from various sources are ingested from external data sources like RDBMS, document data stores, and EDWs. The Hadoop cluster processes, transforms, and analyzes this data using a variety of tools like Pig and Hive. These results may then, in turn, be imported to external data stores for future reference.

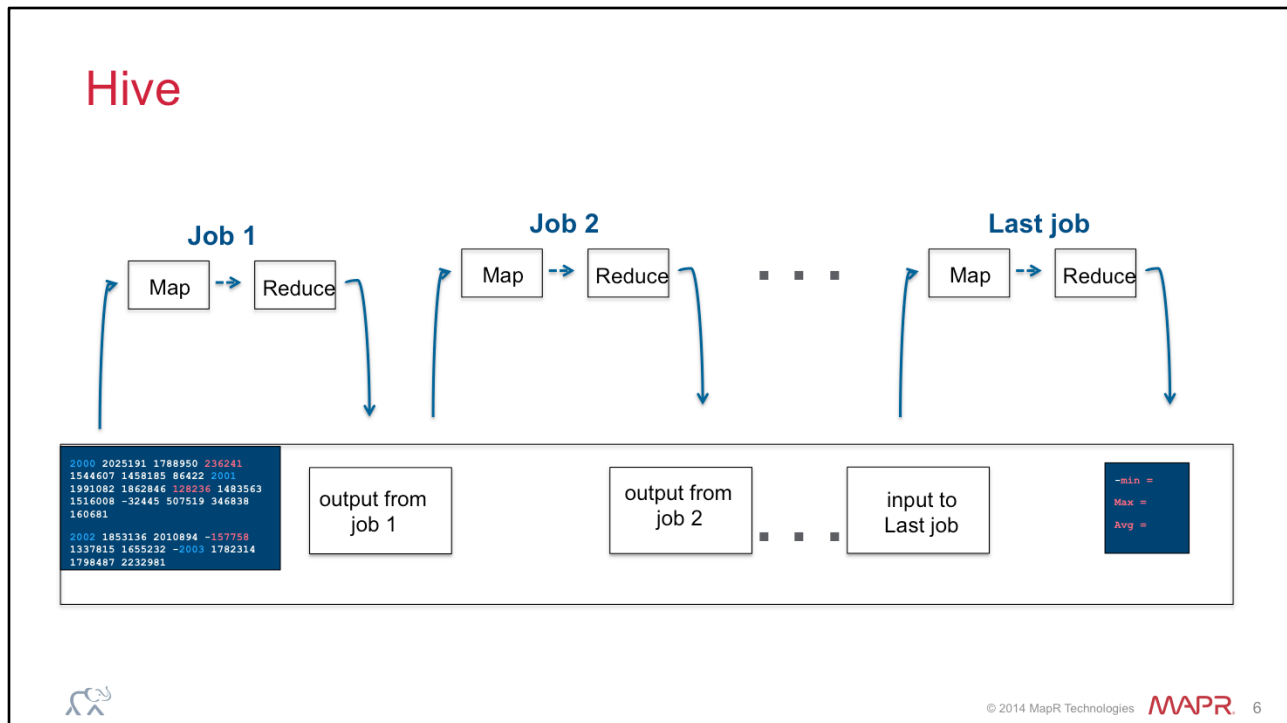


Flume



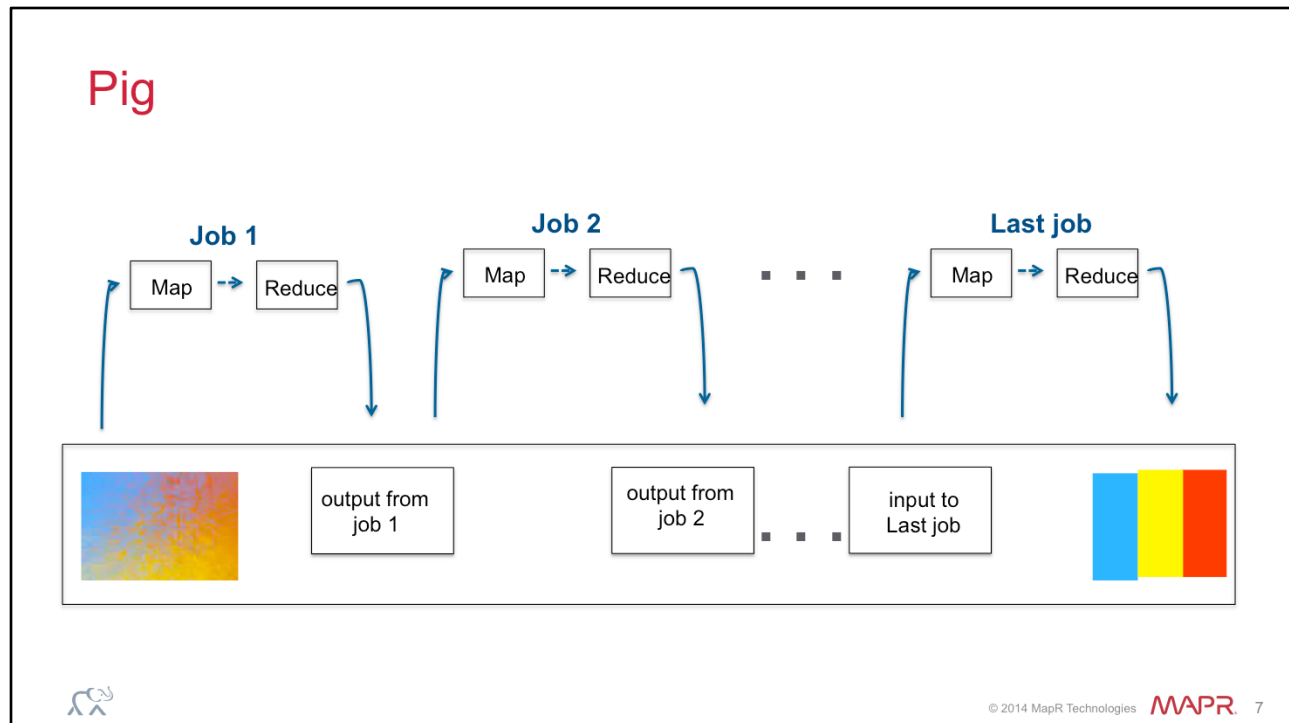
Data comes in all shapes and sizes. Furthermore, some data is like a river or stream – it just keeps flowing in. Examples of streaming data are shown in the slide above, including Syslog logs, Web server logs, Facebook data, and Twitter feeds.





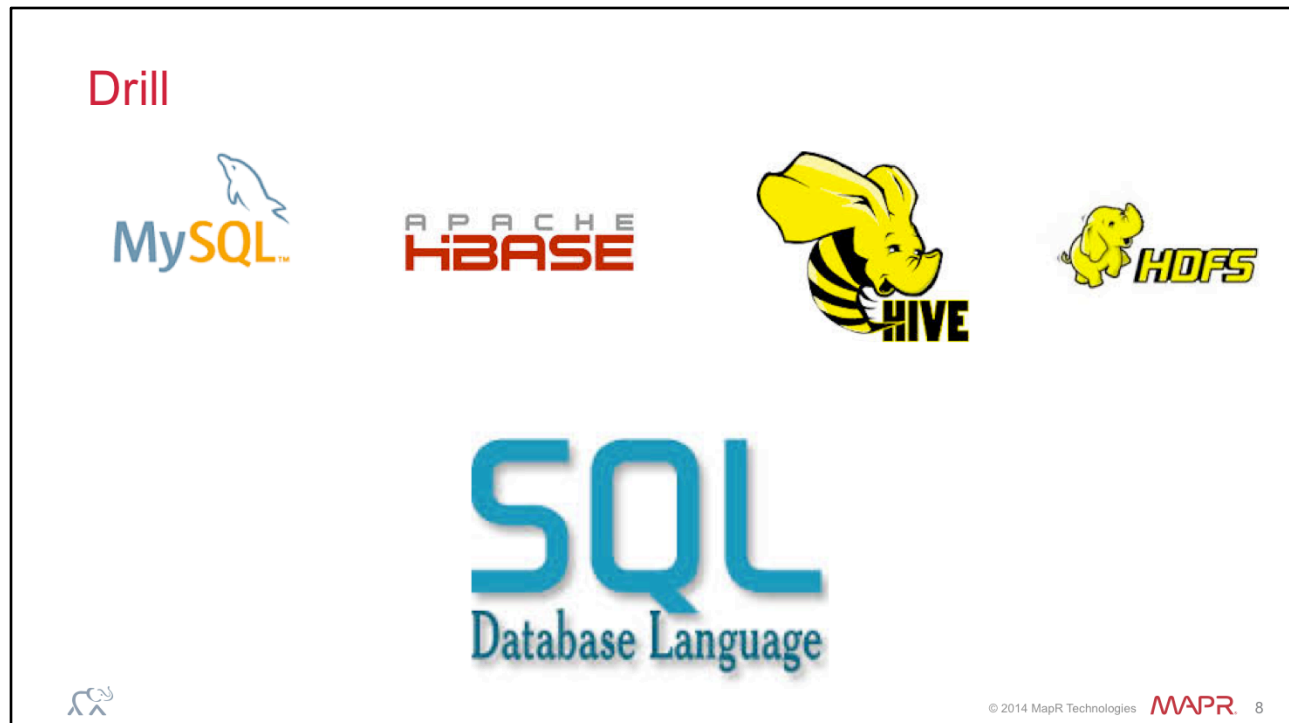
The core mechanism for analyzing data is using MapReduce. Usually, MapReduce jobs must be written as a series of jobs (where each job performs a small set of transformations or calculations). The output from the first job becomes input to the second job, and so on.





You could potentially use MapReduce to analyze this data. Often time, MapReduce solutions require multiple jobs to be run in a series, where the output from the first job becomes input to the second job (and so on).





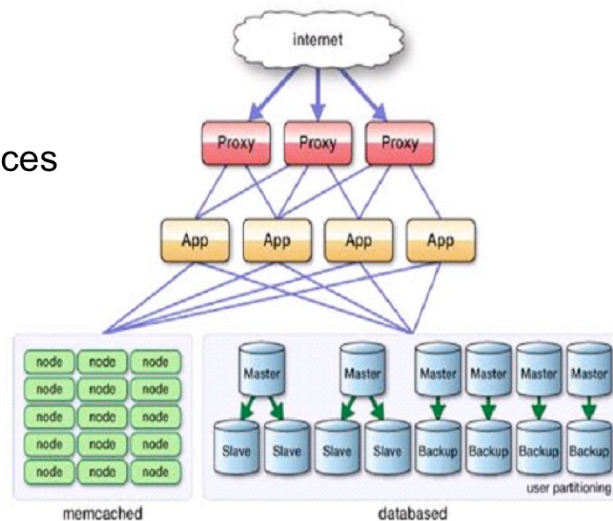
The slide above depicts various data sources that you may wish to analyze at the same time. Each of these data sources has their own interface and set of tools that you can use independently, but how would you analyze all these data sources together in the same query?



HBase

Facebook 2010

- 9000 memcache instances
- 4000 shards MySQL



© 2014 MapR Technologies  9

Back in 2010, Facebook faced the glass ceiling for the practical capacity of an RDBMS solution.

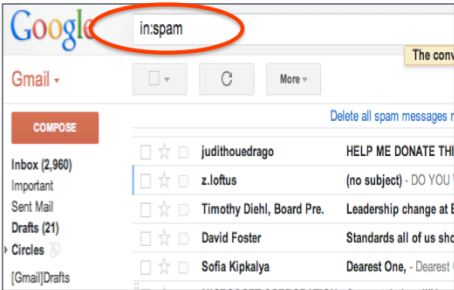
Facebook paired complex sharding and caching to MySQL. Facebook split its MySQL database into ~ 4,000 shards and 9,000 instances of memcached in order to handle the site's massive data volume. This became very difficult to maintain and scale and now Facebook has moved their messaging to hbase.

<http://gigaom.com/2011/07/07/facebook-trapped-in-mysql-fate-worse-than-death/>

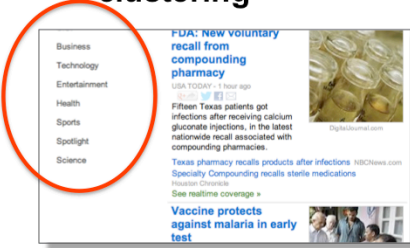


Mahout

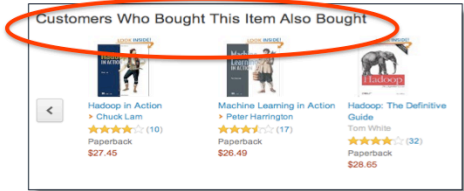
classification



clustering



recommendation



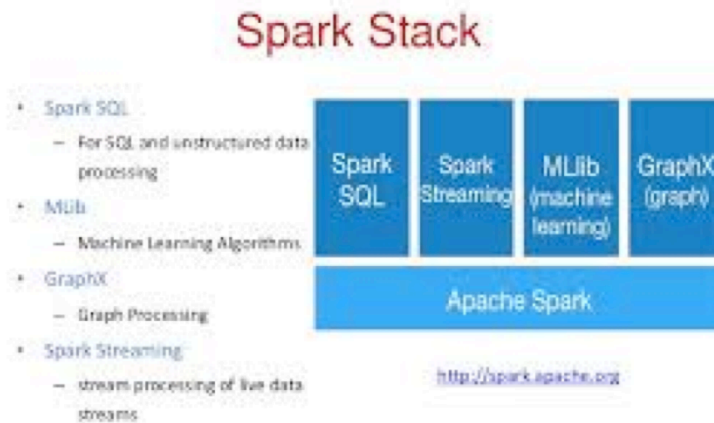
© 2014 MapR Technologies

10

There are several applications of machine learning that are currently being implemented today. Gmail, for examples, uses a ML technique called classification to classify emails as spam or not based on all the data of an email (sender, recipients, subject, message body). Google News uses a technique called clustering to cluster news articles into different categories based on title and content. Amazon uses a ML technique called recommendation (or collaborative filtering) to recommend products to users based on their history and similarity to other users.



Spark



Where Drill provides a SQL interface to multiple underlying data sources and formats, Spark provides an entire processing engine and SQL interface to multiple underlying data sources.

