cloudera

Introduction to Pig

cloudera

In this chapter, you will learn

- What is Pig?
- How Pig is different than Hive

cloudera

What is Pig?

- Data flow language for transforming large data sets
- Originally developed at Yahoo!
- Subproject of Apache Hadoop
  http://hadoop.apache.org/pig

cloudera

Pig is a data flow language for analyzing and transforming large data sets. It has a compiler than translates "Pig Latin" to MapReduce programs. It is an easy to use language with many similarities to SQL and scripting languages.

Pig is an open-source subproject of Apache Hadoop. The official project page is http://hadoop.apache.org/pig. It uses an Apache License, version 2.

Hive vs. Pig

	Hive	Pig
Language	HiveQL (SQL-like)	Pig Latin, a data flow language
Schema	Table definitions that are stored in a metastore	A schema is optionally defined at runtime. Metastore coming soon
Programmatic access	JDBC	PigServer (Java API)

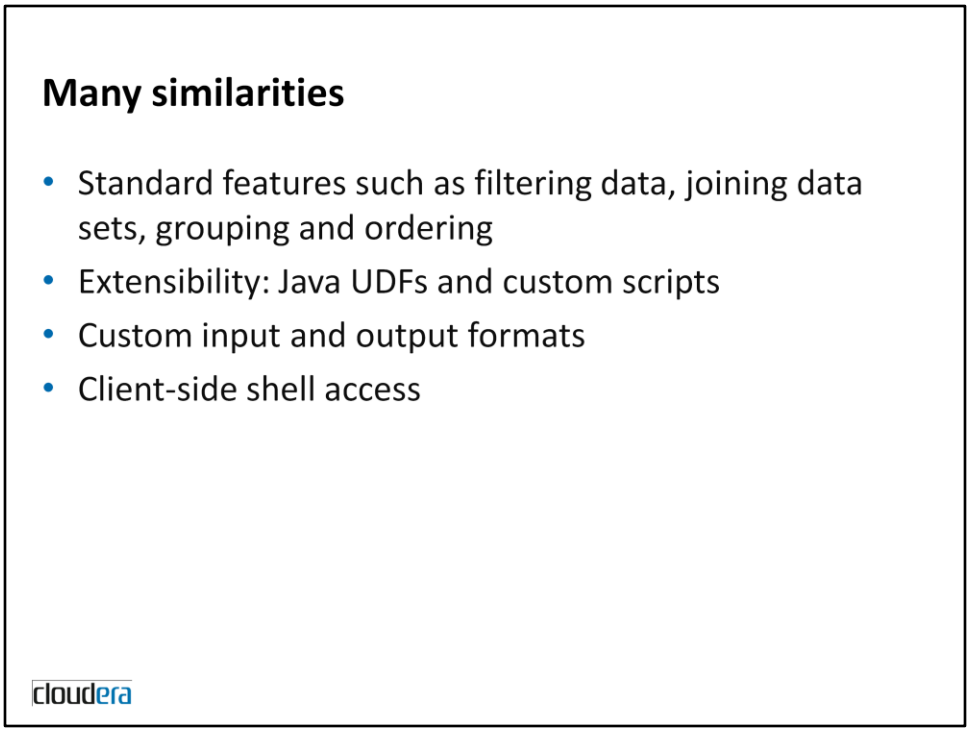
cloudera

Hive and Pig have many similarities, yet take a different approach to analyzing and transforming data sets. Here are some differences:

Language: Hive uses HiveQL which is similar to SQL. HiveQL is a declarative language, meaning that the Hive system has a query planner and optimizer that decides how to execute the command you issue. Pig's language (Pig Latin) allows the programmer to write a set of steps that describe how the data is processed and transformed.

Schema: A significant difference between Hive and Pig is structure of the data. Hive requires the user to define a table that describes the structure of the data. This allows user to share schema definitions (via a centralized metastore). Pig does not have a metastore (yet). Data sets can be described at runtime when the data is processed.

Programmatic access: Both Hive and Pig can be accessed from user programs. Hive uses the standard JDBC (Java Database Connectivity) while Pig uses its own API (`org.apache.pig.PigServer`).



While Hive and Pig have their differences, there are also many similarities.

Both are **higher-level languages** for Hadoop that turn commands into MapReduce programs. Hive and Pig provide for standard data processing features like filtering, joining, grouping and ordering of data.

Hive and Pig are **extensible** by writing user-defined functions (UDFs) in Java and allowing execution of custom scripts (similar to Hadoop Streaming).

Hive's Serializer/Deserializer capability is analogous to Pig's **custom loaders** which allow the programmer to control the input and output format of data.

Finally, Hive and Pig both have a convenient shell for writing commands interactively.

In this chapter, you have learned

- What is Pig?
- How Pig is different than Hive

cloudera