Hadoop Ecosystem Tour

Overview

- Higher-level languages
- Associated projects
- Additional tools

You Say, “tomato…”

Google calls it:	Hadoop equivalent:
MapReduce	Hadoop MapReduce
GFS	HDFS
Sawzall	Hive, Pig
BigTable	HBase
Chubby	ZooKeeper

Pig

- Data-flow oriented language
  - "Pig latin"
  - Datatypes include sets, associative arrays, tuples
  - High-level language for routing data, allows easy integration of Java for complex tasks
- Developed at Yahoo!

Pig Examples

```
named_events FOREACH events_by_time GENERATE
  $1 as event, $2 as hour, $3 as minute;
noon_events = FILTER named_events BY hour =
  '12';
distinct_events = DISTINCT noon_events;



FILTER raw_table BY
  com.cloudera.CustomFilter($1, $2)
```

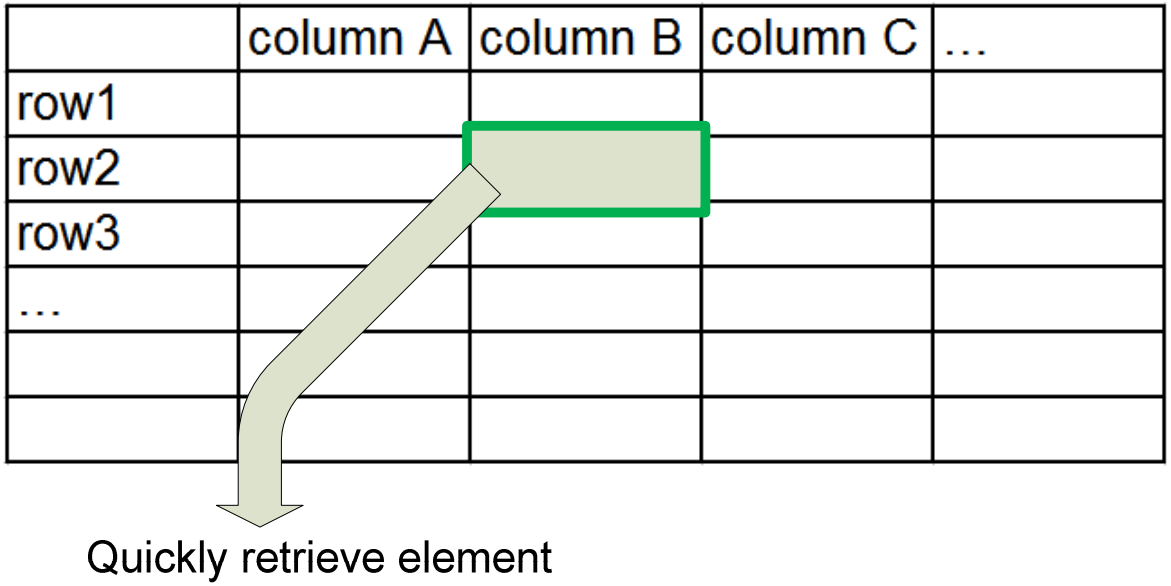
Hive

- SQL-based data warehousing app
  - Feature set is similar to Pig
  - Language is more strictly SQL-esque
- Supports SELECT, JOIN, GROUP BY, etc.
- Features for analyzing very large data sets
  - Partition columns
  - Sampling
  - Buckets
- Developed at Facebook

HBase

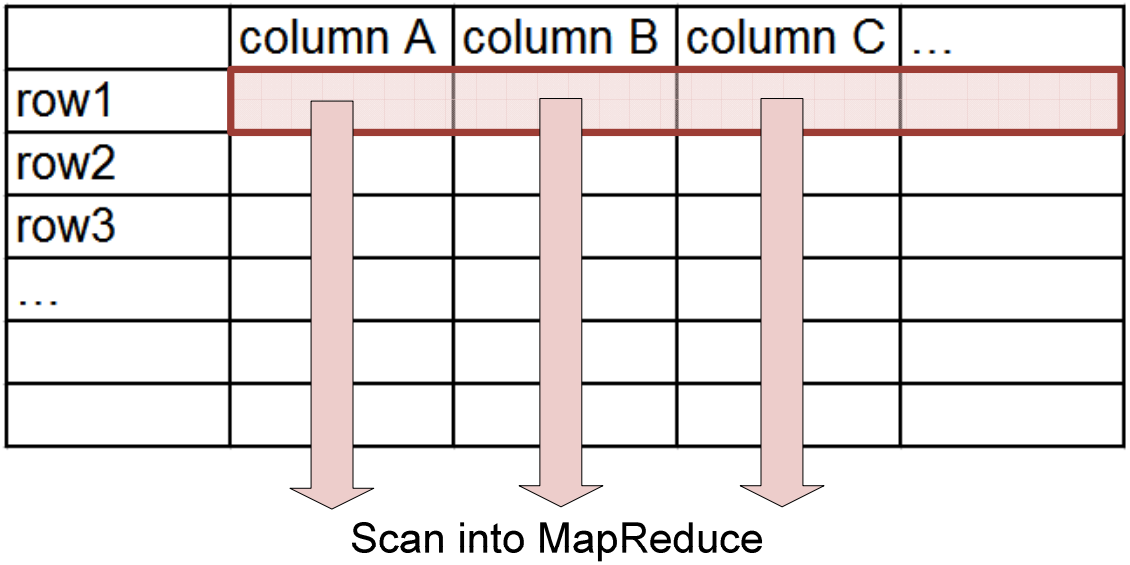
- Column-store database
  - Based on design of Google BigTable
  - Provides interactive access to information
- Holds extremely large datasets (multi-TB)
- Constrained access model
  - (key, val) lookup
  - Limited transactions (only one row)

Fast single-element access



- High-speed lookup of individual (row, column)
- Selects data needed by online applications

HBase as a MapReduce input



- Each row is an input record to MapReduce
- MapReduce jobs can sort/search/index/query data in bulk

ZooKeeper

- Distributed consensus engine
- Provides well-defined concurrent access semantics:
  - Leader election
  - Service discovery
  - Distributed locking / mutual exclusion
  - Message board / mailboxes

fuse-dfs

- Allows mounting of HDFS volumes via Linux FUSE filesystem
  - Does not imply HDFS can be used for general-purpose filesystem
  - Does allow easy integration with other systems for data import/export

Pipes, Streaming

- Multi-language connector libraries for MapReduce
  - Write native-code MapReduce in C++
  - Write MapReduce passes in arbitrary scripting languages

Integrates with other systems

- EC2 launch scripts – run Hadoop in the cloud
  - …with S3-based filesystem implementation
- Ganglia – distributed monitoring

Native-code reimplementations

- Hypertable – Native-code BigTable
- KosmosFS – Native-code GFS-alike

- Not in widespread production use, but worth keeping an eye on

Some more projects…

- Chukwa – Hadoop log aggregation
- Scribe – More general log aggregation
- Mahout – Machine learning library
- Cassandra – Column store database on a P2P backend
- Dumbo – Python library for streaming


cloudera


BY SA