What is Apache Drill

- Inspired by Google Dremel
- Interactive analysis framework of large-scale datasets
- Supports a nested data model
- Plug-able data sources
- Standard SQL 2003 support

Disclaimer: some slides come from http://bit.ly/1c6djHT
Architecture

- Interfaces: REST, CLI, Native API, JDBC, ODBC
- Processing: Query Planner, Execution Engine, Storage Engine
- Data Sources: Cassandra, RDBMS, MongoDB, HBase, HDFS, Hadoop

Architecture

- Each Node: Drillbit – maximize data locality
- Co-ordination, query planning, execution, etc, are distributed
- By default Drillbits hold all roles
- Any node can act as endpoint for a query
- Zookeeper for ephemeral cluster membership info
- Distributed cache (Hazelcast) for metadata, locality information, etc.
Query Execution

- Source query \(\rightarrow\) Parser API
- Custom operators, UDF \(\rightarrow\) Logical Plan
- Serving tree, CF, topology \(\rightarrow\) Physical Plan/Optimizer
- Data sources & formats \(\rightarrow\) Scanner API
Status

- Under development, alpha version
- Current available features
  - Logical Plan
  - Reference interpreter
  - Basic SQL parser
  - Basic demo

Thanks