Using Hadoop as a SQL Data Warehouse (Apache HAWQ)

Lei Chang

Pivotal Inc.
lchang@pivotal.io

FOSDEM'16

Pivotal

- **Introduction**
- **HAWQ 2.0 new features**
- **How to contribute**

Pivotal

HAWQ

Native SQL-on-Hadoop Engine

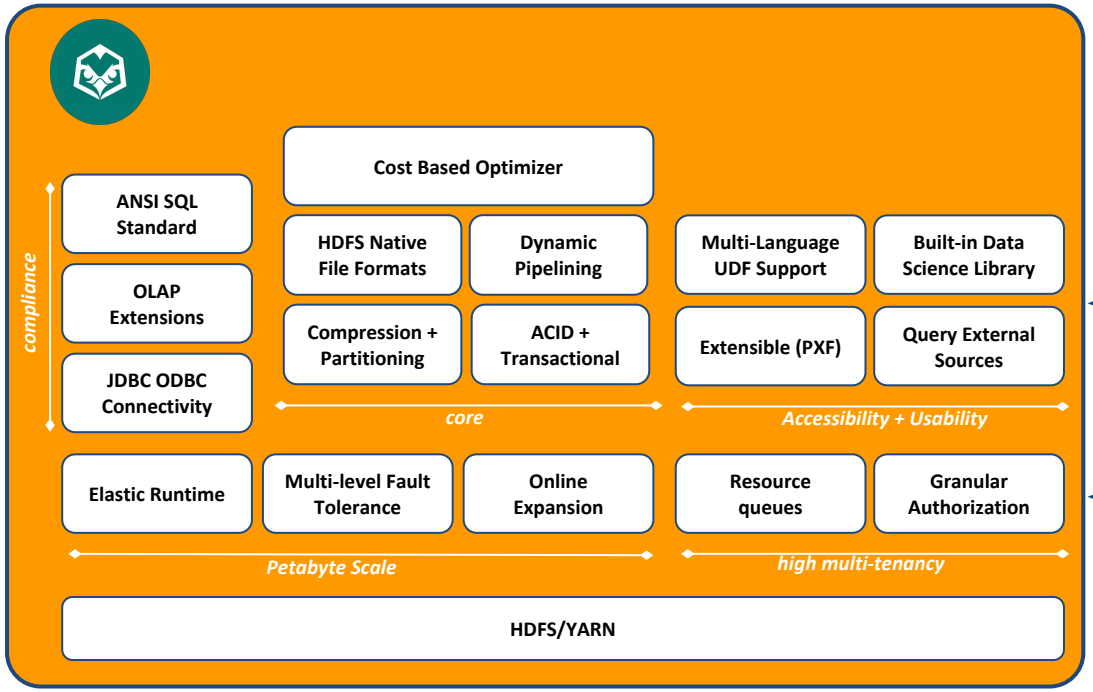
Pivotal

HAWQ Main Features

Apache HAWQ

- Leverage Existing Skills & Tools
- Easily Integrate with Other Tools

- Well Integrated with Hadoop Ecosystem

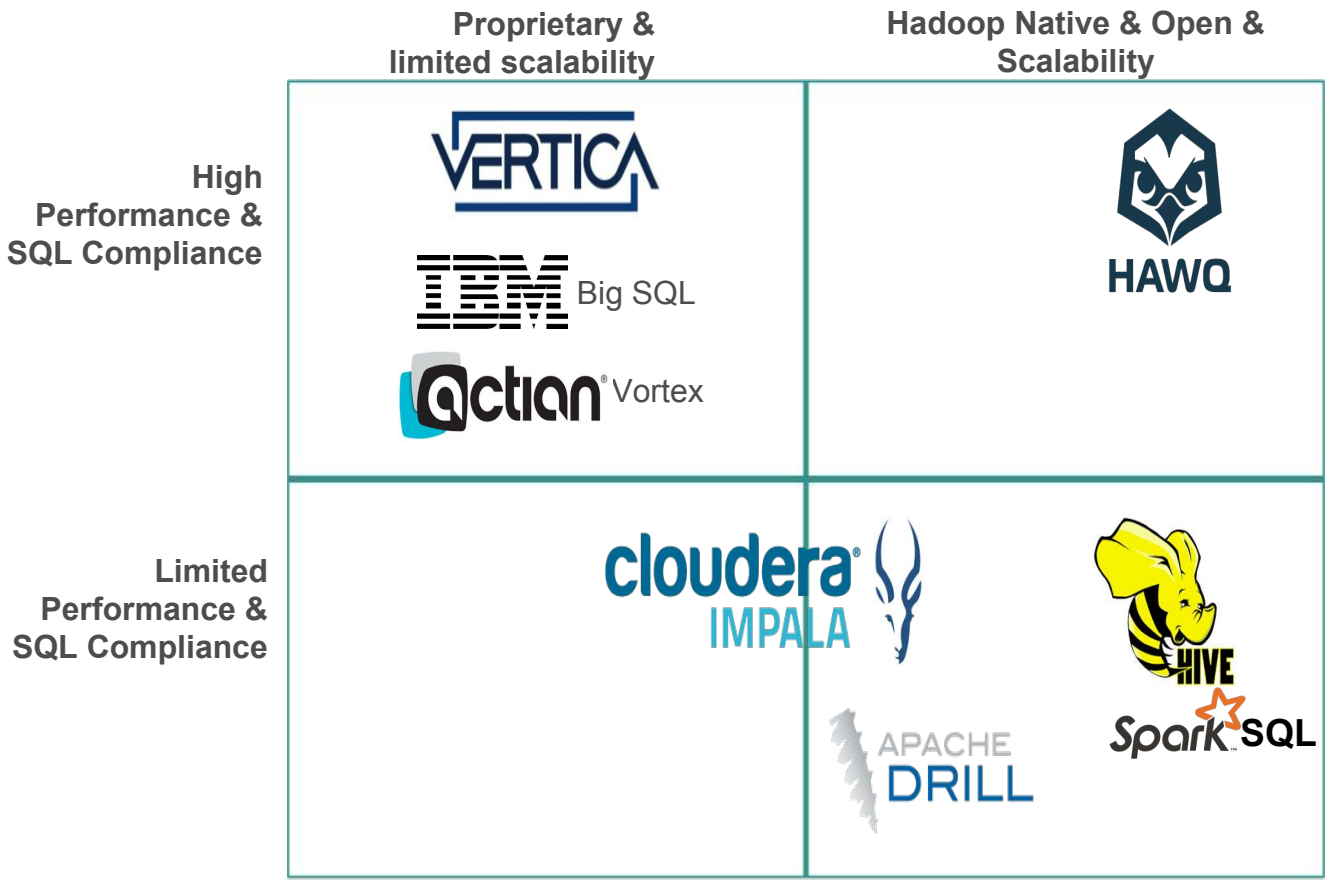


- Discover New Relationships
- Enable Data Science
- Analyze External Sources
- Query All Data Types!

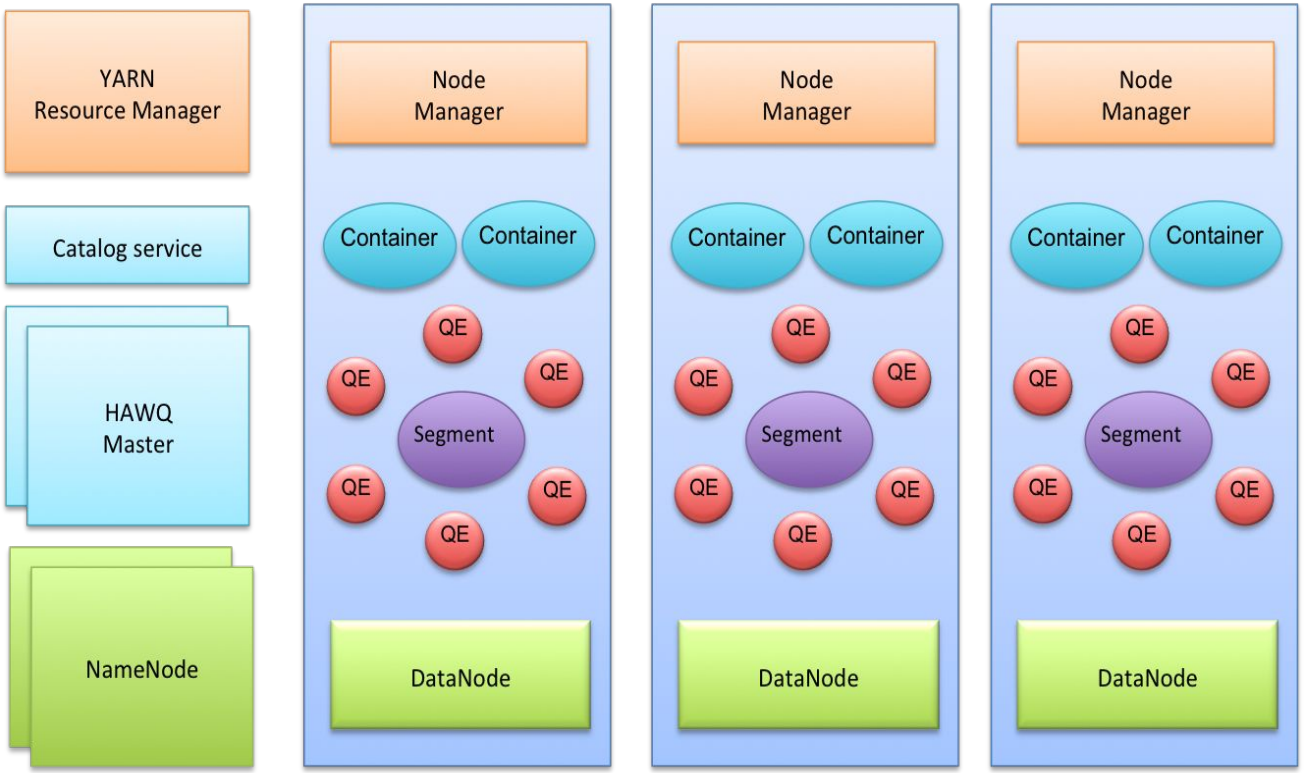
- Manage Multiple Workloads
- Petabyte Scale Analytics
- Sub-second Performance

Pivotal

HAWQ OSS Competitive Positioning

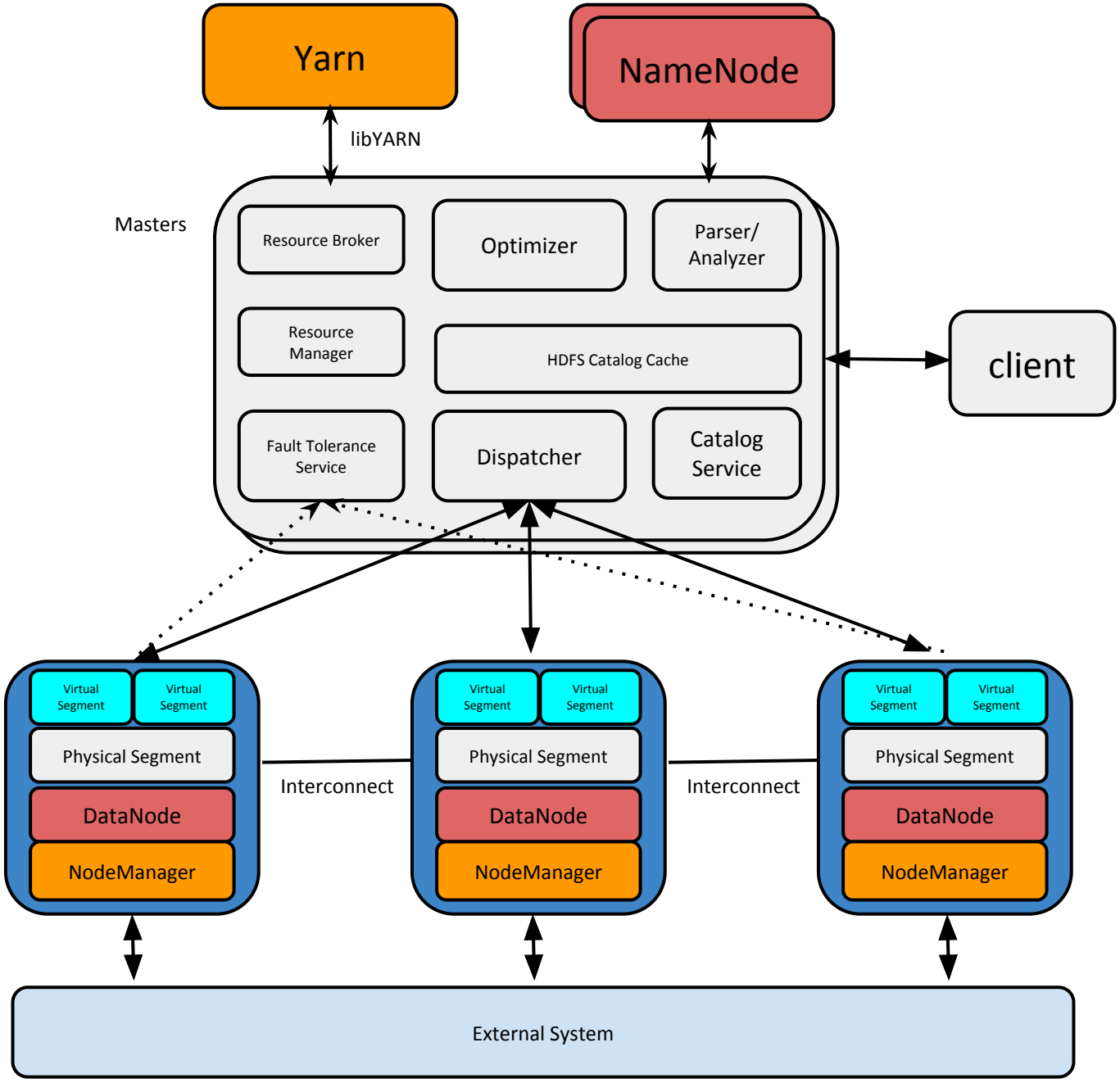


HAWQ components

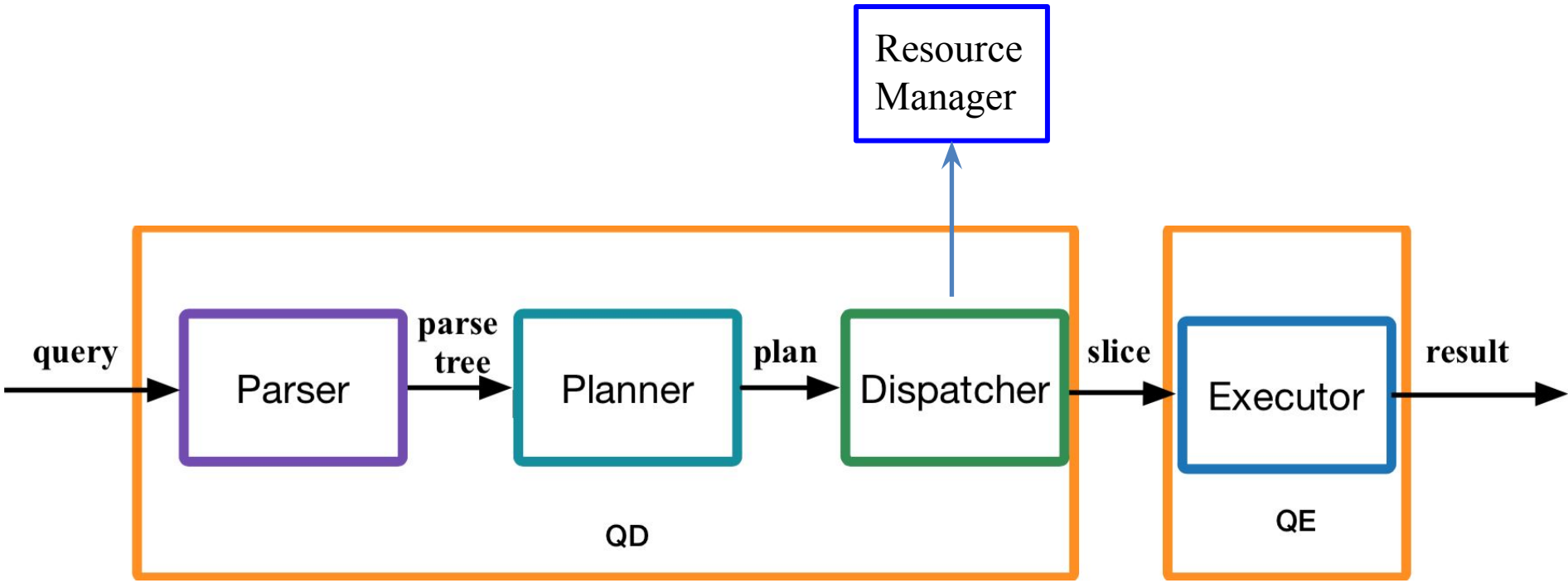


Pivotal

HAWQ Architecture



Basic query execution flow



HAWQ status

2011: Prototype

2012: HAWQ Alpha

2013/03: HAWQ 1.0

Architecture changes for a Hadoopy system

2013~2014: HAWQ 1.x

HAWQ 1.1, HAWQ 1.2, HAWQ 1.3…

2015: HAWQ 2.0 Beta & Apache incubating

http://hawq.incubator.apache.org

2016 Q1/early Q2: HAWQ 2.0 GA

HAWQ 2.0 New Features (In Beta)

- **Elastic execution runtime**
  - One physical segment per node
  - Multiple virtual segments can be started on each node
  - Queries can run on subsets of nodes
- **Resource management**
  - Three layer resource management
  - Global-YARN/Query/Operator
  - Hierarchical resource queues
  - YARN integration
- **Dynamic expansion**
  - fast and without redistribution
- **New dispatcher**
- **New fault tolerance service**
  - Heartbeat and on-demand probe
- **Per table directory**
  - Ease integration with external systems
  - Multi-tenancy
- **Block level storage**
  - AO & Parquet
- **HDFS catalog cache**
  - Accelerate data locality compute
- **New management tools**
  - Consolidate all management tools
- **HCatalog integration**

Pivotal

How to Contribute

Pivotal

Contributing to HAWQ

- Documentation
- Wiki
- Bug reports
- Bug fixes
- Features

- Website: http://hawq.incubator.apache.org/
- Wiki: https://cwiki.apache.org/confluence/display/HAWQ
- Repo: https://github.com/apache/incubator-hawq.git
- JIRA: https://issues.apache.org/jira/browse/HAWQ
- Mailing lists: dev/user@hawq.incubator.apache.org

Code contribution process

- Start a JIRA
- Fork a github repo: https://github.com/apache/incubator-hawq.git
- Clone your repo to local
- Add the github repo as “upstream”
- Create a feature branch and commit your code
- Start a pull request for code review

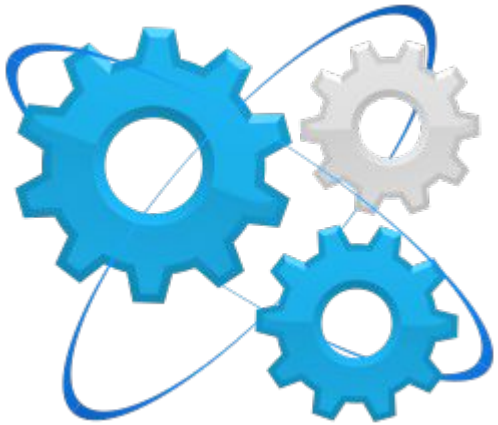
Details:https://cwiki.apache.org/confluence/display/HAWQ/Contributing+to+HAWQ

Interesting areas to contribute

- Indexes
- Update & delete on non-heap tables
- Snapshot
- Geo-replication
- Integrate with ecosystems
- Build on different platforms
  - MacOS, Ubuntu, SUSE

Build & Setup

- Build: https://cwiki.apache.org/confluence/display/HAWQ/Build+and+Install
  - Option 1: Use pre-built docker image:
  - Option 2: Build dependencies by yourself
  - Option 3: “Yum Install” dependencies
- Setup & Run
  - HDFS (required)
  - YARN (Optional)
  - HAWQ Init/start/stop cluster
  - psql -d postgres

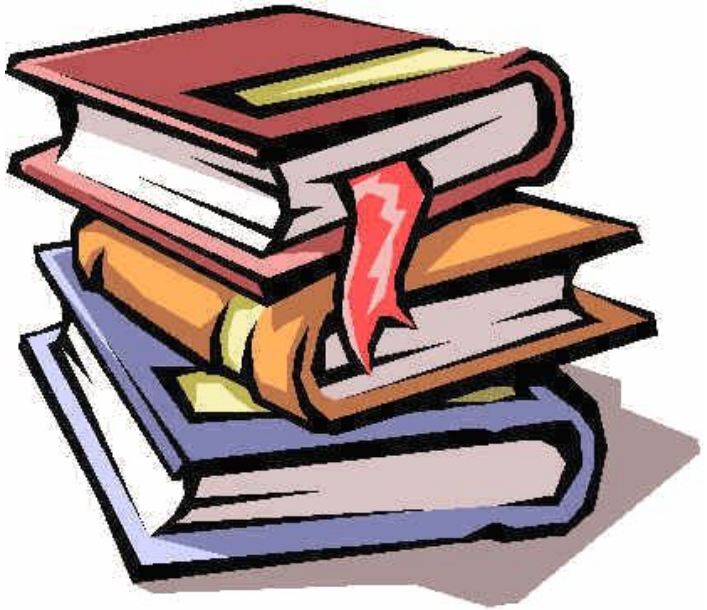


References

- HAWQ website:
  - http://hawq.incubator.apache.org
  - http://pivotal.io/big-data/pivotal-hawq
- HAWQ publications
  - Lei Chang et al: HAWQ: a massively parallel processing SQL engine in hadoop. SIGMOD Conference 2014: 1223-1234
  - Mohamed A. Soliman et al: Orca: a modular query optimizer architecture for big data. SIGMOD Conference 2014: 337-348
  - Lyublena Antova et al: Optimizing queries over partitioned tables in MPP systems. SIGMOD Conference 2014: 373-384
  - Amr El-Helw et al: Optimization of Common Table Expressions in MPP Database Systems. PVLDB 8(12): 1704-1715 (2015)

Summary

- Introduction
- HAWQ 2.0 new features
- How to contribute



Pivotal

Pivotal