Presto: SQL done faster

Strata, London 2017

Wojciech Biela
Lukasz Osipiuk

Teradata Center for Hadoop
About Presenters

Wojciech Biela
Engineering Manager @ Teradata Labs, Center for Hadoop
Wojciech.Biela@Teradata.com

Łukasz Osipiuk
Software Engineer @ Teradata Labs, Center for Hadoop
Lukasz.Osipiuk@Teradata.com
What is Presto?

- **100% open source distributed SQL query engine**
  - Originally developed by Facebook

- **Key requirements:**
  - Performance & Scale
  - Cross platform query capability, not only SQL on Hadoop
  - Hadoop distro agnostic
  - ANSI SQL

- **Apache licensed, hosted on GitHub**
  - Certified distro & support from Teradata
Presto Users

See more at https://github.com/prestodb/presto/wiki/Presto-Users
Presto Users

Multiple clusters (1000s of nodes total)
300PB in HDFS, MySQL, and Raptor
1000s users, 10-100s concurrent queries
Presto uses - Facebook - Data warehouse

- Hive + HDFS + ORC
- multiple clusters
- Thousands of users, 300PB, 1000s nodes
- PBs of data scanned, O(100k) queries every day
- 100s of concurrent queries
Presto uses - Facebook - User facing

- Usage: eg. reporting backend for ad campaign analytics
- Sharded MySQL
- relatively small data (10’s to 100’s of TBs)
- 0.1-5 seconds latency
- Support for data updates
- highly available (different DCs)
- 10-15 way joins
Presto uses - Facebook - various

- Usage: eg. critical dashboards; analytics backend for A/B testing
- Raptor
- medium data (10’s to 100’s of TBs)
- seconds to minutes latency
- predictable performance
- 5-15 minutes load latency
- 100s concurrent queries
250+ nodes on AWS
40+ PB stored in S3 (Parquet)
Over 650 users with 6K+ queries daily
Netflix data pipeline

- Suro / Kafka
- Ursula
- TVs
- mobile
- laptop
- Cassandra
- Aegisthus
- events
- dimensions
- Amazon S3
- TD
- Netflix data pipeline
- presto
Presto Users

200+ nodes (2 dedicated clusters)
25K+ & 3K+ queries daily
Presto Users

200+ nodes on-premise
Parquet nested data
Presto Users

120+ nodes in AWS
2PB is S3 and 200+ users
supported by Teradata
Presto architecture

- Client
- Coordinator
  - Metadata API
  - Data location API
  - Parser/analyzer
  - Planner
  - Scheduler
- Worker
  - Data stream API
- Worker
  - Data stream API

Pluggable components highlighted in red.
Presto architecture

- **Plugin architecture**
  - Core Presto
    - SQL Parser
    - planner/Optimizer
    - execution engine
  - Plugins
    - connectors - data+metadata
    - extensions
      - user defined functions
      - event listeners
      - access controllers
Core Features - simple deployment

**Simple deployment**
- self contained (RPM/tar.gz)
- worker auto discovery
- trivial dependencies
  - just recent JVM
- single-port network communication
  - easy firewall/network setup
- even easier with presto-admin
Core Features - hardware agnostic

- **Infrastructure agnostic**
  - on premise (appliance or commodity clusters)
  - VM (OpenStack, etc.)
  - cloud (Amazon, Azure, etc)
    - pure EC2
    - EMR
    - AWS Athena (pay-as-you-go)
Connectors - Supported data sources

Coordinator

Parser/ analyzer

Metadata API

Scheduler

Data stream API

Planner

Data location API

Worker

HDFS / S3

NoSQL

DBMS

Custom

..
Connectors - Hive

● Table metadata read from Hive catalog
● Multiple filesystems
  ○ HDFS
  ○ S3
● Supported file formats
  ○ ORC (optimized reader, optimized writer in flight)
  ○ RCFile (optimized reader, optimized writer)
  ○ Parquet (optimized reader)
  ○ Avro
  ○ SequenceFile
  ○ Text
Connectors - Hive

- Client
- Presto Coord
- Hive Metastore
- HDFS NameNode
- HDFS DataNode

Connections:
- Hive metadata
- HDFS block locations
- Data
Deployment strategies - Hadoop

Presto + HDFS
Deployment strategies - Hadoop

Separate cluster
Deployment strategies - Hadoop

Shared machine

- Presto Worker
- HDFS DataNode
- Presto Worker
- HDFS DataNode
- Presto Worker
- HDFS DataNode
Deployment strategies - Hadoop

Mixed - rack local
Connectors - Raptor

● Native Presto columnar store
  ○ mostly used internally in FB

● Performance characteristics
  ○ Large data sets (trillions of rows, 100s TB)
  ○ Seconds to minutes latency
  ○ Predictable performance
  ○ 5-15 minute load latency
  ○ 10s of concurrent queries

● Transactional loads

● Data organization based on temporal column

● Data bucketing for collocated joins

● Under heavy development and poorly documented
Connectors - Raptor

- **Two tier storage**
  - columnar store on flash drives on workers
  - slower backup storage

- **Tables sharding**

- **Shard replication**
Connectors - Raptor

- Client
- Presto Coord
- Raptor metadata (MySQL)
- Presto Worker
- SSD
- SSD
- SSD
- SSD
- SSD
- SSD
- SSD
- SSD

- Shards
- Metadata
- Data

Backup storage tier
Connectors - supported data sources

● Other data sources
  ○ SQL databases (reads non parallel)
    ■ MySQL
    ■ PostgreSQL
    ■ Microsoft SQL Server
  ○ Cassandra
  ○ Kafka
  ○ Redis
  ○ MongoDB
  ○ ElasticSearch
  ○ Accumulo
Core Features - SQL support

- ANSI SQL support
  - all standard data types
  - complex subqueries support
    - support for correlated subqueries (some limitations)
Core Features - SQL support

- **Structural types**
  - map
  - array
  - JSON

- **Lambda expressions support**
  - SELECT transform (ARRAY['dog', 'whale'], x -> length(x))
    ■ [3, 5]
  - SELECT reduce(ARRAY [5, 20, 50], 0, (s, x) -> s + x, s -> s)
    ■ 75
Core Features - SQL support

● All standard DDL/DML is supported
  ○ CREATE TABLE / CREATE TABLE AS
  ○ connector specific extensions supported via WITH clause
  ○ DROP TABLE
  ○ INSERT
  ○ DELETE
  ○ GRANT / REVOKE

● Set of supported features depends on connector
  ○ richest support for Hive and Raptor connectors
Core Features - Connectivity

- **CLI**
- **Closed source JDBC/ODBC drivers (available free of charge)**
  - full JDBC ODBC spec compliance
  - Kerberos authentication
  - LDAP authentication
  - sponsored by Teradata
- **Open source JDBC driver**
  - does not support Java 6
  - limited support to authentication
- **Open source ODBC driver**
  - not maintained
- **Language specific bindings (RPresto, python, ruby, ...)**
BI Tools certifications

Tableau

Power BI

Information Builders

Zoomdata

MicroStrategy

Qlik
Core Features - performance

● MPP style in-memory execution
● Vectorized data processing
● Highly tuned Java
  ○ Query to ByteCode compilation
  ○ Memory efficient structures - Minimize GC
  ○ Careful inner loop implementation
● Multi-threaded execution keeps CPU busy
  ○ Focus on being versatile. Support both single query at a time and highly concurrent workloads.
Core Features - performance

- **Optimizer**
  - currently rule based
  - in flight cost based

- **Exploit statistics provided by connectors**
  - Hive catalog statistics

- **Cost based decisions for current optimizer**
  - Join type selection
  - Join reordering in progress
Core Features - security

- User authentication (CLI/ODBC/JDBC)
  - Basic
  - Kerberos
  - LDAP
- Pluggable user authorization schemes (access controllers)
- Connector level authorization
  - E.g. grants information stored in Hive catalog
- Support for kerberized HDFS/Hive metastore
- SSL on the wire
  - client to Presto
  - between Presto nodes
Teradata QueryGrid™ - Multi-System Analytics

Entry Points
- Teradata Database
- Aster Analytics
- Presto Hadoop

Targets
- Integrated Data Warehouses
- Multi-Genre Advanced Analytics™
- Multiple Hadoop SQL Query Engines and Distributions
- 3rd Party Relational DBs
- Non-Relational DBs

Presto Connectors
- MySQL
- PostgreSQL
- Redis
- Apache Cassandra
- Apache Kafka
- Presto API

Teradata Certified

Community Supported
Questions?

www.teradata.com/presto