CSE 599C - Spring 2017 - University of Washington

Vertica Interactive Tutorial

By Helga, Cong, Tony and Maaz

Tutorial Layout

1. About Vertica: Pricing, Features & Performance

2. Using Vertica: Setup, Data Ingestion, Running Queries

3. Optimizing Vertica: Tuning the database

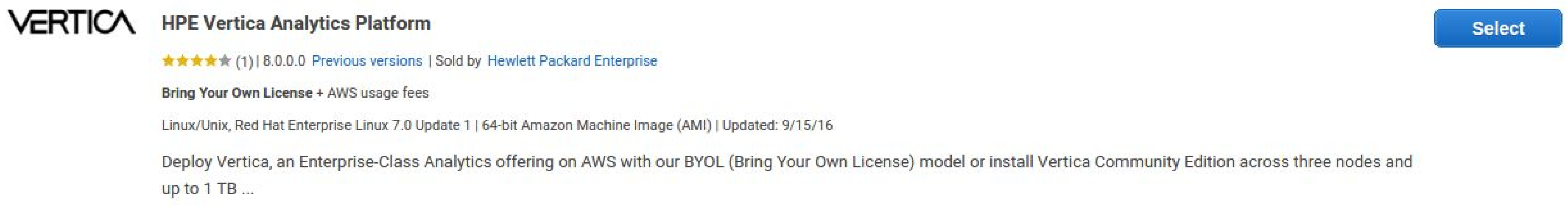
4. Data Analytics with Vertica: A case study with time-series analysis

Quick Recap: What is Vertica?

- Distributed Big Data Analytics database by HP.
- Designed to handle terabytes / petabytes of data.
- Column-oriented storage design.
- Runs on major Linux distributions (Ubuntu, Debian, Suse, RHEL)
- Relational Database.
- Supports SQL (Many interfaces: vSQL, JDBC / ODBC drivers etc)

About the product

- Offered on-premise, in the cloud and directly on top of Hadoop
- Free Community Edition License: 3 Nodes and up to 1TB
- Amazon AMI available for running Vertica on AWS
- Community vs Enterprise license details here.



HPE Vertica Analytics Platform

★★★★★ (1) | 8.0.0.0 Previous versions | Sold by Hewlett Packard Enterprise

Bring Your Own License + AWS usage fees

Linux/Unix, Red Hat Enterprise Linux 7.0 Update 1 | 64-bit Amazon Machine Image (AMI) | Updated: 9/15/16

Deploy Vertica, an Enterprise-Class Analytics offering on AWS with our BYOL (Bring Your Own License) model or install Vertica Community Edition across three nodes and up to 1 TB ...

Recap: Key Concepts

- Column based storage
  - Improved I/O performance
- Projections
  - Optimize frequent queries
- Clustering
  - MPP
  - Data segmented across nodes
  - Fault tolerance
  - Elastic Scaling

Cool Features

- Vertica is extensible
- UDFs can be created using R, C++ or Java
    - R for scalar and transform functions
    - Java for analytic and load functions in “fenced” mode.
    - C++ for all functions in “fenced” or “unfenced” mode

Cool Features

- Provides machine learning functions for in-database analysis!
- Can store machine learning models.
- Can perform data preparation and predictive tasks
  - K-means
  - Linear Regression
  - Logistic Regression

Cool Features

- Built-in analytical functions:
    - Time series interpolation
    - Event-based sessionization
    - Pattern matching
    - Geo-spatial analysis

Cool Features

- Workload Analyzer
    - Analyzes information in system tables
    - Makes tuning recommendations

Is Vertica right for you?

- CRUD vs Analytics
- Performance Comparison: Postgres vs Vertica [1]
    - PostgreSQL 9.2
    - Vertica Analytic Database v7.2
    - Flights data: ~36m records
    - Single Node Virtual Machine

Query 1: Count records

```sql
SELECT count(*) FROM flight_fact;
```

execution	count(*)	PostgreSQL	Vertica	% of PostgreSQL response time
1	35874731	30951ms	44ms	0.14%
2	35874731	30989ms	53ms	0.17%
3	35874731	29973ms	36ms	0.12%

Query 2: Number of flights by airport

```sql
SELECT airport_origin_id, count(*)
FROM flight_fact
GROUP BY airport_origin_id;
```

execution	PostgreSQL	Vertica	% of PostgreSQL response time
1	28100ms	883ms	3.14%
2	27904ms	869ms	3.11%
3	28228ms	818ms	2.90%

Query 3: Airports with most departures

```
SELECT a.airport_fullname_name, count(*)
FROM flight_fact f JOIN airport_dim a ON f.airport_origin_id = a.airport_id
GROUP BY a.airport_fullname_name
ORDER BY count(*) DESC LIMIT 20
```

execution	PostgreSQL	Vertica	% of PostgreSQL response time
1	28548ms	6253ms	21.16 %
2	27237ms	4966ms	18.23%
3	26390ms	5103ms	19.34%

Query 4: Busiest days of the year

```
SELECT d.year, d.fullday, count(*)
FROM flight_fact f JOIN date_dim d ON f.date_id = d.date_id
GROUP BY d.year, d.fullday
ORDER BY d.year, count(*) DESC;
```

execution	PostgreSQL	Vertica	% of PostgreSQL response time
1	46200ms	8912ms	19.29%
2	52165ms	7892ms	15.13%
3	51785ms	7103ms	13.72%

Demo!