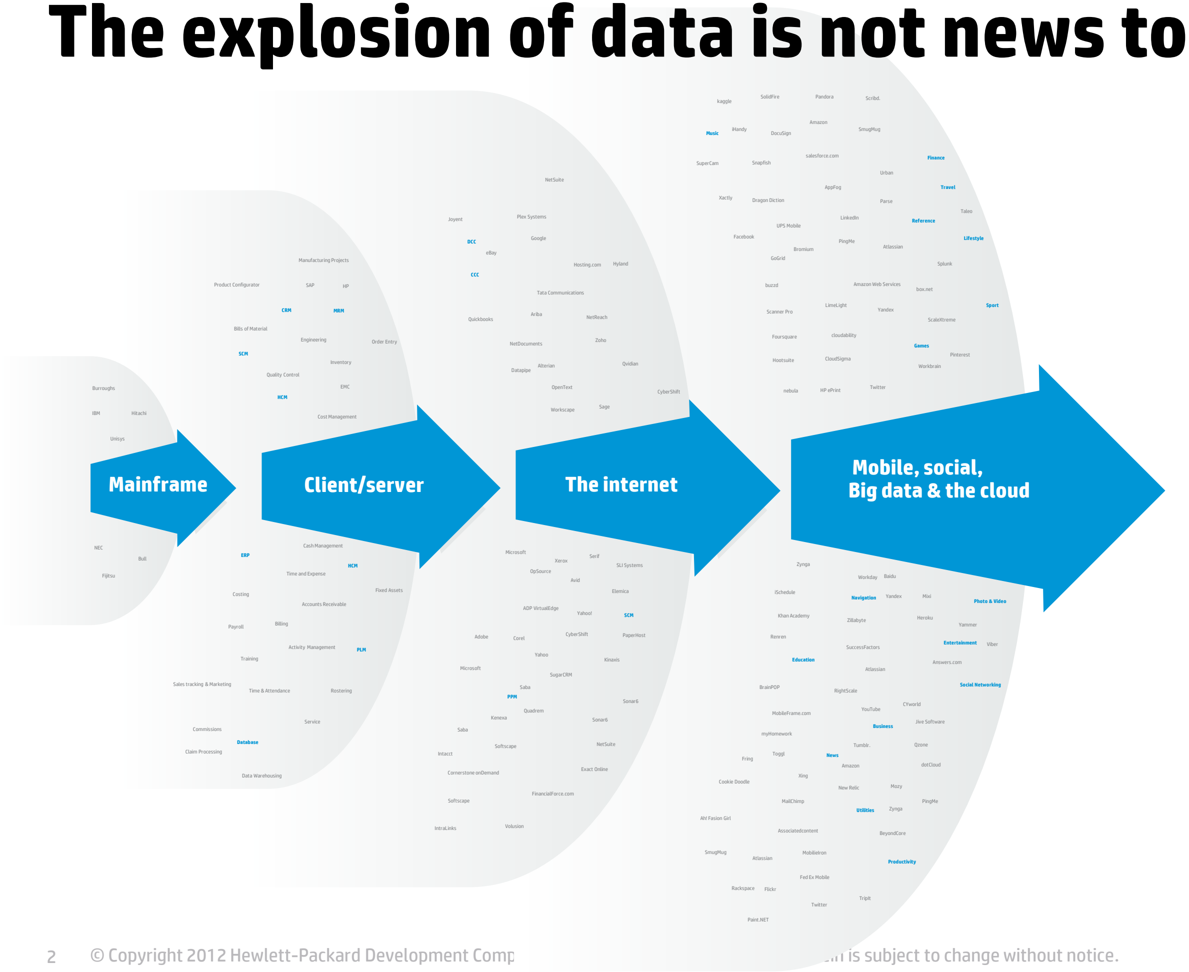
Real-time Collaborative Analytics with Vertica Analytic Platform

Stephen Walkauskas,

Architect, Data Management, Vertica

C-Big October 14, 2012

The explosion of data is not news to anyone ...



Mainframe

Client/server

The internet

Mobile, social,
Big data & the cloud

Every 60 seconds

98,000
tweets

23,148
+apps downloaded

400,710
ads requests

208,333
minutes Angry Birds played

2000
lyrics played on Tunewiki

Today, data analysis is slow, painful and costly

Imagine a world where you can have a conversation with your data!

Vertica makes this a reality with real time analytics !!!

VERTICA
An HP Company

Introducing Vertica

- SQL Database for Real-time Analytics
- Runs on x86 hardware
- MPP Columnar Architecture – scales to PBs!
- Reduced footprint via Advanced Compression
- Extensible analytics capabilities
- Easy to setup and use
- Elastic - grow/shrink as needed
- Extensive Ecosystem of analytic tools

Speed

+

Scale

+

Simplicity

Proven by 2500+ Customers Worldwide

zynga
MUSIC TELEVISION
EMPIRIX
Bell
Admeld
Honeywell
meteor GAMES
accenture
ca technologies
SOE SONY ONLINE ENTERTAINMENT
Disney Interactive Media Group
GSN
Pizza Hut
comcast
Vonage
AERA
R M S Risk Management Solutions
MEDEANALYTICS MEASURE. MANAGE. LEAD.
XM SIRIUS SATELLITE RADIO

- Promotional Testing
- Claims Analyses
- Patient Records Analyses
- Clinical data Analyses
- Fraud Monitoring
- Financial tracking
- Tick data back-testing

- Behavior Analytics
- Click Stream Analyses
- Network Analyses
- Customer Analytics
- Compliance Testing
- Loyalty Analysis
- Campaign Management

5 Building Blocks for Collaborative Analytics

Performance that enables Interactive and Iterative Q&A with the Data

Extensible + ability to share (tools, views, code and data)

Ability to record and replay Analyst “thought-process”

Sand-boxing data to enable ad-hoc and intense experimentation

Ability to dynamically access a variety of data sources

Vertica can be an excellent platform for collaboration!

Performance that enables Interactive and Iterative Q&A with the Data

“Vertica opened doors to analyses that otherwise were either too time-intensive or impossible. A larger team of business managers now have faster, easier access to more information. That knowledge is invaluable in an aggressively competitive market like ours.”

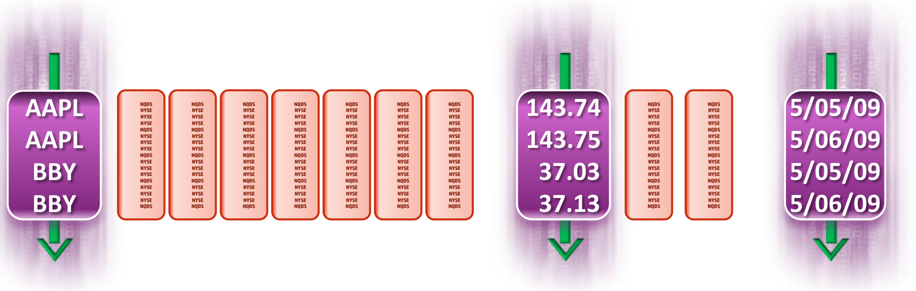
-Brian Harvell, Executive director for network operations, Comcast

Column Store Innovation: Efficiency

Column Store – Column-Based Disk I/O

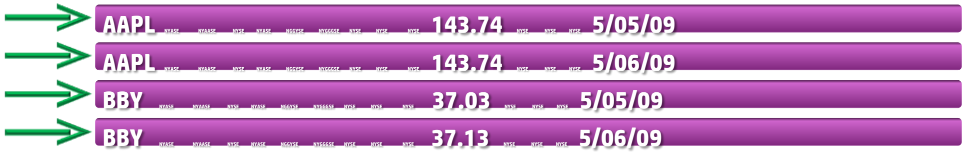
Typical FinServ price per stock for 1 day

Column Store - Reads 3 columns



SELECT AVG(price)
FROM tickstore
WHERE
symbol = ‘AAPL” AND date = ‘5/06/09’

Row Store - Reads all columns



Column Store – Sort and Encode for Speed

Student_ID	Name	Gender	Class	Score	Grade
1256678	Cappiello, Emilia	F	Sophomore	62	D
1254038	Dalal, Alana	F	Senior	92	A
1278858	Orner, Katy	F	Junior	76	C
1230807	Frigo, Avis	M	Senior	64	D
1210466	Stober, Saundra	F	Junior	90	A
1249290	Borba, Milagros	F	Freshman	96	A
1244262	Sosnowski, Hillary	F	Junior	68	D
1252490	Nibert, Emilia	F	Sophomore	59	F
1267170	Popovic, Tanisha	F	Freshman	95	A
1248100	Schreckengost, Max	M	Senior	76	C
1243483	Porcelli, Darren	M	Junior	67	D
1230382	Sinko, Erik	M	Freshman	91	A
1240224	Tarvin, Julio	M	Sophomore	85	B
1222781	Lessig, Elnora	F	Junior	63	D
1231806	Thon, Max	M	Sophomore	82	B
1246648	Trembley, Allyson	F	Junior	100	A

Column Store – Sort and Encode for Speed

Gender	Class	Grade	Score	Name	Student_ID
F	Sophomore	D	62	Cappiello, Emilia	1256678
F	Senior	A	92	Dalal, Alana	1254038
F	Junior	C	76	Orner, Katy	1278858
M	Senior	D	64	Frigo, Avis	1230807
F	Junior	A	90	Stober, Saundra	1210466
F	Freshman	A	96	Borba, Milagros	1249290
F	Junior	D	68	Sosnowski, Hillary	1244262
F	Sophomore	F	59	Nibert, Emilia	1252490
F	Freshman	A	95	Popovic, Tanisha	1267170
M	Senior	C	76	Schreckengost, Max	1248100
M	Junior	D	67	Porcelli, Darren	1243483
M	Freshman	A	91	Sinko, Erik	1230382
M	Sophomore	B	85	Tarvin, Julio	1240224
F	Junior	D	63	Lessig, Elnora	1222781
M	Sophomore	B	82	Thon, Max	1231806
F	Junior	A	100	Trembley, Allyson	1246648

Columns used in predicates

Correlated values "indexed" by preceding column

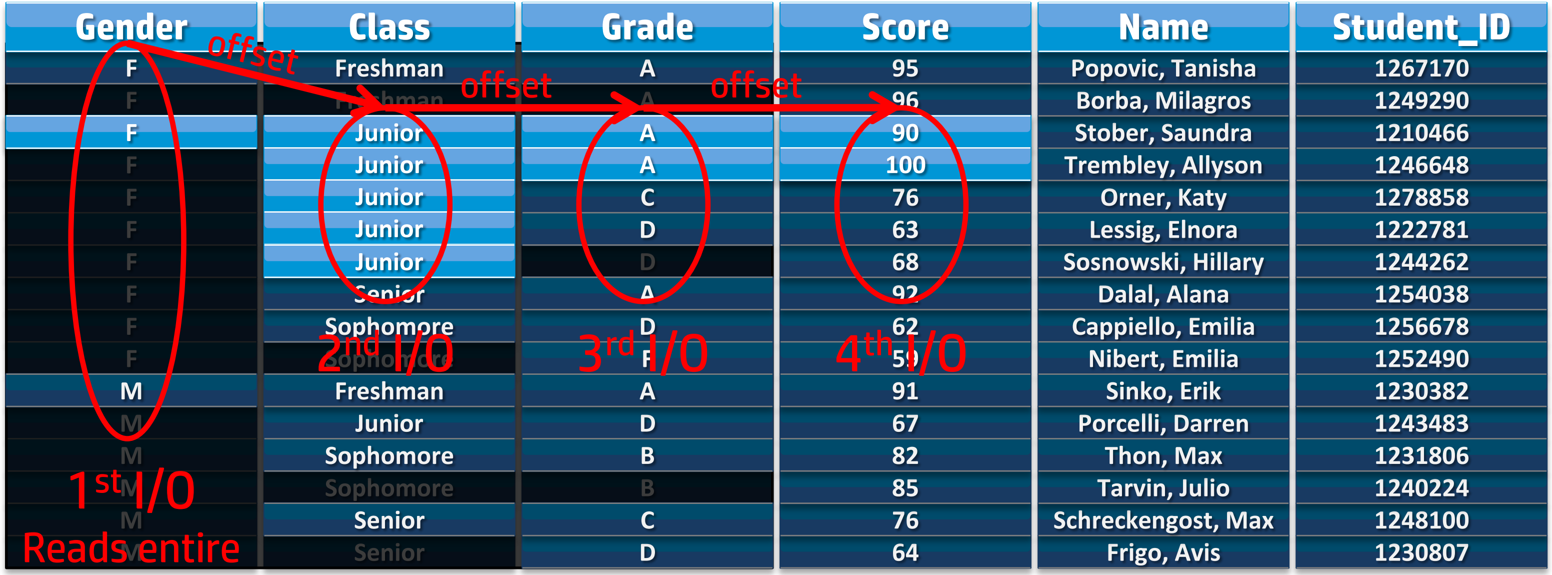
Column Store – Sort and Encode for Speed

Gender	Class	Grade	Score	Name	Student_ID
F	Freshman	A	95	Popovic, Tanisha	1267170
F	Freshman	A	96	Borba, Milagros	1249290
F	Junior	A	90	Stober, Saundra	1210466
F	Junior	A	100	Trembley, Allyson	1246648
F	Junior	C	76	Orner, Katy	1278858
F	Junior	D	63	Lessig, Elnora	1222781
F	Junior	D	68	Sosnowski, Hillary	1244262
F	Senior	A	92	Dalal, Alana	1254038
F	Sophomore	D	62	Cappiello, Emilia	1256678
F	Sophomore	F	59	Nibert, Emilia	1252490
M	Freshman	A	91	Sinko, Erik	1230382
M	Junior	D	67	Porcelli, Darren	1243483
M	Sophomore	B	82	Thon, Max	1231806
M	Sophomore	B	85	Tarvin, Julio	1240224
M	Senior	C	76	Schreckengost, Max	1248100
M	Senior	D	64	Frigo, Avis	1230807

Columns used in predicates

Correlated values "indexed" by preceding column

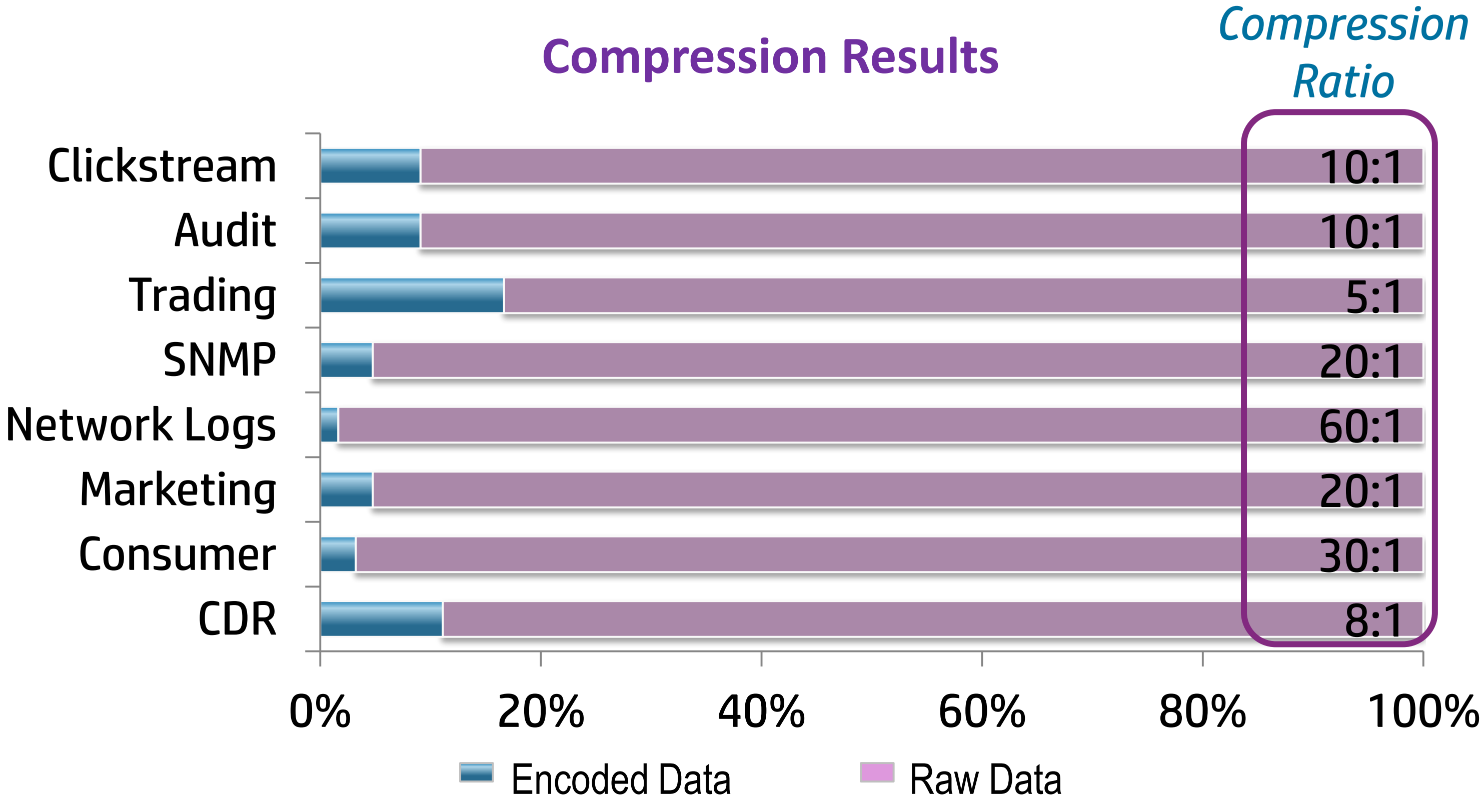
Column Store – Sort and Encode for Speed



Gender	Class	Grade	Score	Name	Student_ID
F	Freshman	A	95	Popovic, Tanisha	1267170
F	Freshman	A	96	Borba, Milagros	1249290
F	Junior	A	90	Stober, Saundra	1210466
F	Junior	A	100	Trembley, Allyson	1246648
F	Junior	C	76	Orner, Katy	1278858
F	Junior	D	63	Lessig, Elnora	1222781
F	Junior	D	68	Sosnowski, Hillary	1244262
F	Senior	A	92	Dalal, Alana	1254038
F	Sophomore	D	62	Cappiello, Emilia	1256678
F	Sophomore	F	59	Nibert, Emilia	1252490
M	Freshman	A	91	Sinko, Erik	1230382
M	Junior	D	67	Porcelli, Darren	1243483
M	Sophomore	B	82	Thon, Max	1231806
M	Sophomore	B	85	Tarvin, Julio	1240224
M	Senior	C	76	Schreckengost, Max	1248100
M	Senior	D	64	Frigo, Avis	1230807

Example query: select avg(Score) from example where Class = ‘Junior’ and Gender = ‘F’ and Grade =

Column Store – Column Based Encoding

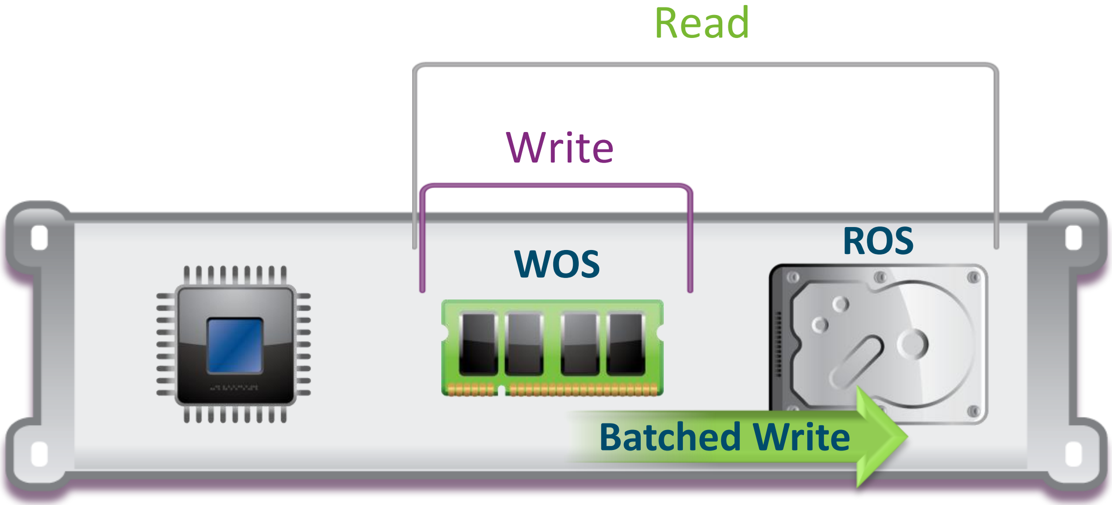


Real-Time Loading and Querying

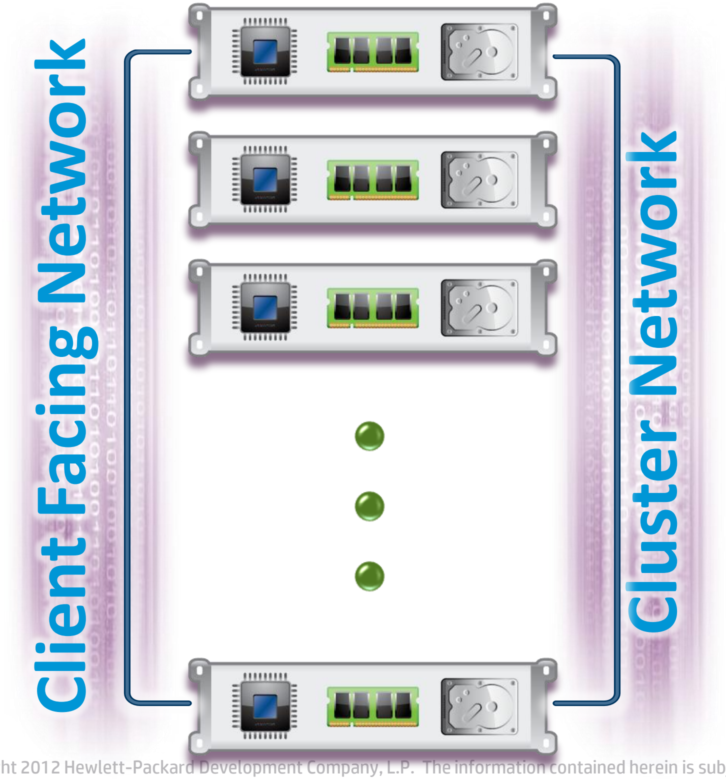
Write-Optimized Store – WOS

Read-Optimized Store – ROS

Tuple Mover – TM



Shared-Nothing, Scale-Out Architecture



Massively Parallel Processing (MPP)

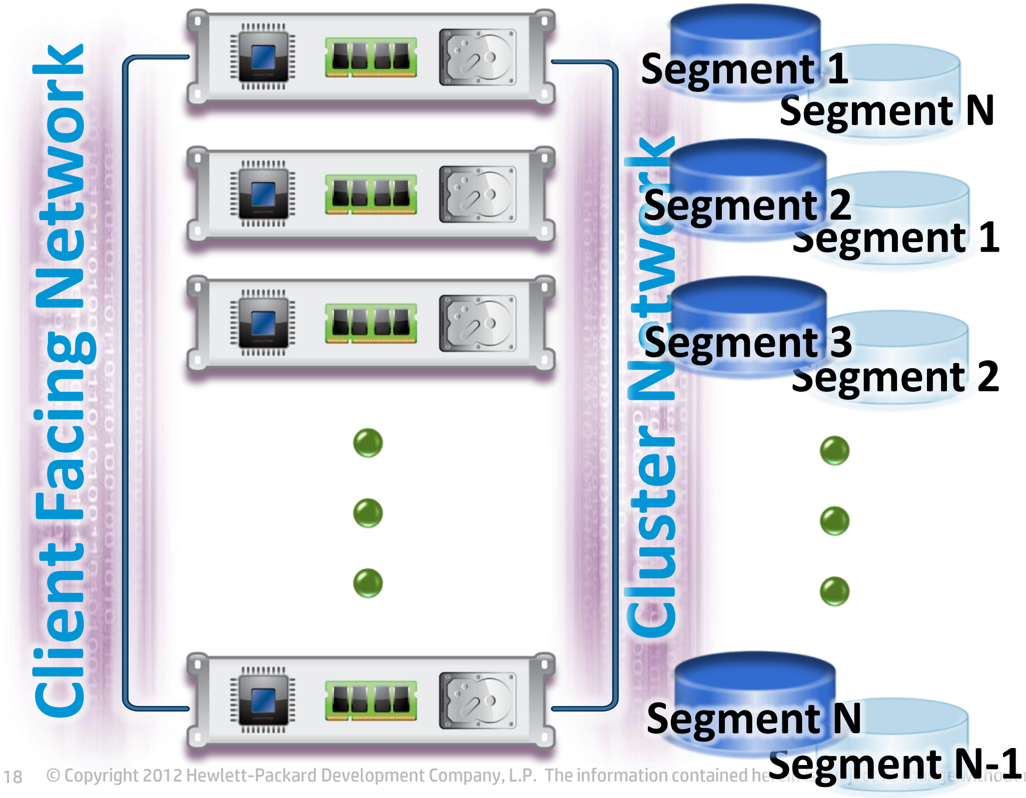
100% peer-to-peer

No specialized nodes

Can query & load to any node

Linear scalability

Revolutionary High Availability



RAID-like functionality within DB

Smart K-safety

Always-on loads & queries

Extensible + Ability to Share

A rich analytic platform with a large set of built-in analytics

Extensibility to develop custom algorithms

Share tools, views, code and data

Ability for users to run analytics defined by someone else via a standard tool

Vertica has a Rich Analytics Platform

SQL

- Window functions
- Graph
- Monte Carlo
- Statistical
- Geospatial

Extended SQL

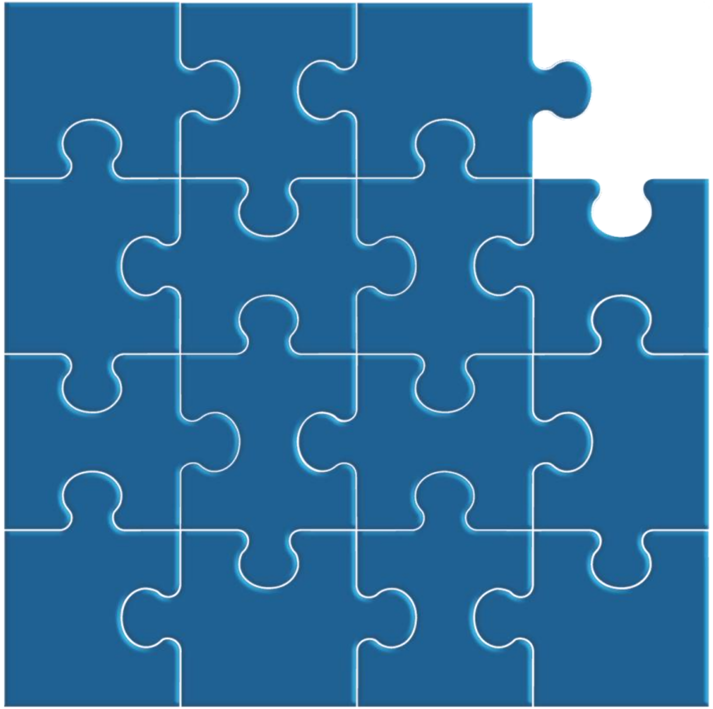
- Sessionization
- Time series
- Pattern matching
- Event series joins

SDKs

- C++
- R
- More to come

Check out: http://www.vertica.com/2011/10/05/being-green-with-data-exhaust

Vertica Analytics Platform SDK



A framework for User-defined Extensions

Languages: C/C++ and R

Simple: concise APIs and examples accelerate deployment

Flexible: operate on Structured and Unstructured data sets, fenced Option for Security

Efficient: In-process, fully parallel

User Community: Github.com

Check out: https://github.com/vertica/Vertica-Extension-Packages

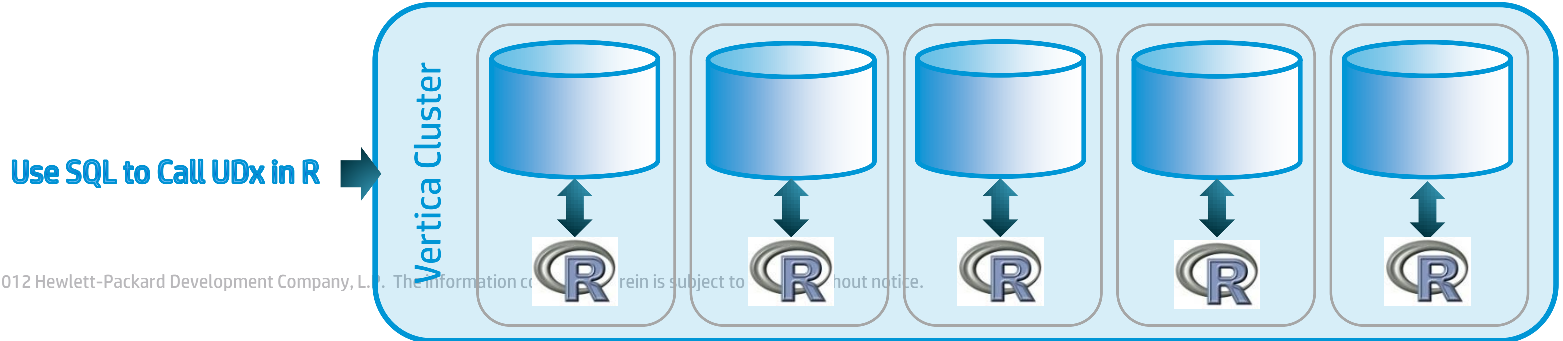
Analytic Extensions in R

What is R?

- Open source language for statistical computing
- Wide range of packages available for advanced data mining and statistical analysis

Advantages of running R from inside Vertica

- Vertica automatically parallelizes the execution of user defined R code
- Optimized data transfer between Vertica and R
- Enable ‘R’ users to benefit from Vertica’s scalable MPP platform
- Enable ‘SQL’ users to benefit from advanced analytics with R



R Analytics Use Cases

Data-Mining Algorithms

K-Means Clustering – Segment customers based on geography, usage patterns, etc

Page Rank – Identify the influencers among my customers / users

K-Nearest Neighbors Classification

Naïve Bayes Classification

Classification and Regression Trees

In-Database Scoring

Financial Services: What is the probability of default for each mortgage in our portfolio?

Sensor data: What is the probability of failure for each of my in-home devices?

Health care: What is the probability that this medical insurance claim is fraudulent?

Record and replay Analyst “thought-process”

Data Collector

Comprehensive tracking of what the database is doing

Automatic and low over-head collection

Includes query logging, performance profiling and so on

Easy SQL access to retrieve data back

Privilege model for sharing / protecting access to activity

Customizable Retention Policy

Sand-boxing data to enable ad-hoc and intense experimentation

Efficient snapshot objects, with COW semantics

Export from one Vertica Cluster to another

SQL command to transfer data subset to another cluster

Source and Target cluster can be different in size, config and physical design

Optimized protocol for data transfer, keeping data compressed when possible

Ability to export to a cluster on EC2

Ability to dynamically access a variety of data sources

User-defined Loads & External Tables

UDL is an extensible adapter API to load data from any source, in any format

External Tables provides ability to make loads “dynamic” i.e. at query time

Connectors available to data sources such as HDFS, Other Databases, etc

Want a conversation with your data? Evaluate Vertica!

Enterprise Edition

- Free 30 day evaluation

Community Edition

- Free Download 1TB, 3 nodes

Check Out Vertica Extensions on Github!

- https://github.com/vertica/Vertica-Extension-Packages

Vertica Community Edition Beta Signup

Get Started.

Get your hands on the Vertica Community Edition to see, firsthand, how quickly and easily you can install, deploy, and start analyzing with Vertica.

Sign Up Now

Sign Up for the Beta Program

First Name

Last Name

Company

Phone

Email

Zip/Postal Code Country

Sign Up

Sign Up for the Beta Program Today

The Vertica Community Edition will be released before the end of the year, with a beta program kicking off very soon. If you would like to be one of the first to test out the Vertica Community Edition, please complete the Sign Up form on the top right of this page and we'll let you know when you can download the software.

Vertica Community Edition / Info Sheet

Download the Vertica Community Edition information sheet

Download the PDF

Community Tools and Resources – Coming Soon!

Vertica Community Edition software and third-party drivers

Ask questions and share your knowledge with the rest of the community.

Testimonials

And especially for Baseball fans!

http://www.vertica.com/2012/09/11/vertica-moneyball-and-r-the-perfect-team/