CITRIX®

Using HPCC Systems for Big Data and More - Because Who Has Time for MapReduce?

John Andleman

October 7, 2014

About Me

I love to architect and build systems that acquire, manage, and use data to solve problems

- Operational
- Analytical
- Real-time
- Big Data
- Data Science

CITRIX®

About Citrix SaaS Division

a market-leading global provider of web collaboration, remote access, data sharing and IT support software as a service.

GoToMeeting

For online meetings

GoToWebinar

For do-it-yourself webinars

GoToTraining

For online training

GoToAssist

for integrated IT support tools

GoToMyPC

for remote access to your Mac or PC

Sharefile

for data sharing and storage

Podio

for social collaboration

OpenVoice

for affordable audio conferencing

CITRIX®

Finding Insights In Big Data

- Structured and semi-structured data
- Data sets from very small to many billions of records
- Hundreds of terabytes of log files
- Thousands of Oracle database tables
- Spreadsheet data
- And more…

CITRIX®

Finding Insights In Big Data – Traditional BI?

- Oracle Data Warehouse
- ROLAP and Data Cubes

- Very expensive licensing and hardware costs
- Does not scale well to very large data sets
- ETL to get data loaded is complicated
- Extracting useful content from log files is complicated
- Limited analytic capabilities

Finding Insights In Big Data – Hadoop?

- It's very powerful, but…

- Why do they have to make it so complicated?!
- MapReduce scales, but it is a giant step backwards in productivity
- Java is a horrible language for data processing; Python is a little better
- Extracting useful content from log files is very complicated
- Much of the Hadoop infrastructure is immature and poorly documented

Finding Insights In Big Data – Hadoop with Pig?

- Much more productive than writing MapReduce code, but…

- The language is very limited
- Where the language has gaps, you end up writing user-defined functions, or worse, going back to writing MapReduce code
- Extracting useful content from log files is still very complicated

Finding Insights In Big Data – HPCC?

- ECL Language is a very mature data processing language
- ECL is a very complete language
- ECL has very powerful pattern matching constructs for extracting useful content from log files – the best I have seen!
- ECL is the best ETL language I have worked with
- HPCC with ECL scales well and is a very productive development environment

CITRIX®

Big Data Projects at Citrix – GoToMeeting and GoToWebinar

- Study product feature usage by:
  - Different customer segments
  - Trial vs. paid customers
  - Retained vs. lost accounts
- Study trial usage patterns of converted vs. non-converted accounts
- Study relationships of various session statistics to customer retention
- Study usage patterns of VoIP vs. dial-up audio in sessions
- Study patterns of session audio problems
- Study to identify fraudulent usage including trial abuse and spam activity

CITRIX®

Big Data Projects at Citrix – HPCC and ECL

- Raw Oracle data was dumped to CSV files for ingestion into HPCC – this turned out to be faster than doing any data crunching in Oracle
- HPCC for ETL jobs ran MUCH faster than Oracle, even when HPCC was run on much smaller hardware
- HPCC is a much more capable and productive ETL tool than anything else I have used. I even prefer it for data that is not "big".
- ECL has excellent support for analytics, especially on very large data sets that would choke most other analytic tools

CITRIX

Work better. Live better.