Unify Analytics
Making Production Data Accessible for Data Science at Scale

Paige Roberts
Vertica Open Source Relations Manager
Analytics Use Cases are Everywhere

Companies that enable our data-driven world use both BI and data science.

- **IoT Analytics**
- **AdTech Analytics**
- **COVID-19 Tracking**
- **Customer Analytics**
- **Energy Trading**
- **Security Analytics**

- **Network Optimization**
- **Predictive Maintenance**
- **Energy Optimization**
- **Smart Agriculture**
- **Software Optimization**
- **Clickstream Analytics**

© 2020 All rights reserved.
Data Warehouse Architecture

**Analytical Database**

**BATCH**
- ERP
- CRM
- Billing

Transactional data
- Files
- Application Data
- Customer
- Operational
- Financial

Extract, Transform, Load (ETL)

**Analytical Database**
- Reporting
- Ad Hoc Queries

**Business Intelligence**

**SQL**

**BUSINESS INTELLIGENCE REPORTS, VISUALIZATION**

© 2020 All rights reserved.
Data Warehouse Strengths

**DATA WAREHOUSE**

- Reporting/Business Intelligence
- High Performance
- High Concurrency
- Reliability
- Security
- Governance
- SQL
Data Warehouse Weaknesses

**DATA WAREHOUSE**
- Reporting/Business Intelligence
- High Performance
- High Concurrency
- Reliability
- Security
- Governance
- SQL

**WEAKNESSES**
- Expensive to scale
- Structured data only
- ETL caused data to be stale
- Business intelligence only
- Can’t handle streaming data
## Data Lake Architecture

### LOW LATENCY
- Application data
- Web clicks
- Logs
- Sensors
- Operational metrics
- User tracking
- Geo-location

### BATCH
- Contextual data
- Weather
- Geo
- Files

- Transactional data
- Application Data
- OLTP/ODS

---

**Mass Storage (Data Lake)**

- HDFS

**Data Prep**

- ELT with Transformation in data lake

**Prepared data in new format**

**SQL-Like Query Engine**

- SQL-Like Query Engine

- Python
- Jupyter
- R

---

**Contextual data**

- Weather
- Geo
- Files

- Object Storage

**Machine Learning**

- DATA SCIENCE VISUALIZATION, APPLICATIONS

---

© 2020 All rights reserved
Data Lake Strengths

- Machine Learning/Data Science
- Unlimited Scale
- Streaming Data
- Semi-Structured Data (JSON, AVRO, ...)
- Complex Data Types (Maps, Structs, Arrays)
- Schema on Read
- Python, R, Jupyter
Data Lake Weaknesses

WEAKNESSES
- Slow performance
- Poor concurrency
- Complex to build / maintain
- Immature reliability, security, governance
- Difficulty operationalizing ML

DATA LAKE
- Machine Learning/ Data Science
- Unlimited Scale
- Streaming Data
- Semi-Structured Data (JSON, AVRO, ...)
- Complex Data Types (Maps, Structs, Arrays)
- Schema on Read
- Python, R, Jupyter
Cooperative Architecture Strengths

DATA WAREHOUSE
- Reporting/Business Intelligence
- High Performance
- High Concurrency
- Reliability
- Security
- Governance
- SQL

DATA LAKE
- Machine Learning/Data Science
- Unlimited Scale
- Streaming Data
- Semi-Structured Data (JSON, AVRO, ...)
- Complex Data Types (Maps, Structs, Arrays)
- Schema on Read
- Python, R, Jupyter
Cooperative Architecture

LOW LATENCY
Application data
- Web clicks
- Logs
- Sensors
- Operational metrics
- User tracking
- Geo-location

Distributed
- Pub/Sub
- kafka

Stream Processing

Mass Storage
(Data Lake)
- HDFS
- Parquet
- Object Storage
- Machine Learning

Histotical data

AND / OR
Fast ELT with Transformation in data lake

ELT, Data Prep

SQL

BUSINESS INTELLIGENCE + DATA SCIENCE

BATCH
Contextual data
- Weather
- Geo

Files

Distributed
Analytical Database

Transaction data
Application Data
OLTP/ODS

LOW LATENCY
Application data
- Web clicks
- Logs
- Sensors
- Operational metrics
- User tracking
- Geo-location

Distributed
- Pub/Sub
- kafka

Stream Processing

Mass Storage
(Data Lake)
- HDFS
- Parquet
- Object Storage
- Machine Learning

Histotical data

AND / OR
Fast ELT with Transformation in data lake

ELT, Data Prep

SQL

BUSINESS INTELLIGENCE + DATA SCIENCE

BATCH
Contextual data
- Weather
- Geo

Files

Distributed
Analytical Database

Transaction data
Application Data
OLTP/ODS
**LOW LATENCY**
- Planting and harvest equipment
- Weather stations, probes, satellite imagery
- Application data - clickstreams

**BATCH**
- Bayer research trials
- Climate research farms (CRF)
- Climate research partners (CRP)
- FieldView data
- Environmental data
- Platform partner data
- Sales data
- Marketing campaigns

---

**Mass Storage (Data Lake)**

**Separate cluster for data Ingest, ETL**

---

**Distributed Analytical Database**

**Machine Learning**

**SQL**

**Data Science + Business Intelligence**

---

**LOW LATENCY**

- **Mass Storage (Data Lake)**
  - Amazon RDS
  - Machine Learning

**BATCH**

- **Distributed Analytical Database**
  - Machine Learning

---

**LOW LATENCY**

- **Mass Storage (Data Lake)**
  - Amazon Kinesis
  - Amazon RDS

**BATCH**

- **Distributed Analytical Database**
  - Machine Learning

---

**LOW LATENCY**

- **Mass Storage (Data Lake)**
  - Amazon Kinesis

**BATCH**

- **Distributed Analytical Database**
  - Machine Learning

---

**LOW LATENCY**

- **Mass Storage (Data Lake)**
  - Amazon Kinesis

**BATCH**

- **Distributed Analytical Database**
  - Machine Learning
Cooperative architecture weaknesses

**DATA WAREHOUSE**
- Reporting/Business Intelligence
- High Performance
- High Concurrency
- Reliability
- Security
- Governance
- SQL
- BI Visualization tools

**DATA LAKE**
- Machine Learning/Data Science
- Unlimited Scale
- Streaming Data
- Semi-Structured Data (JSON, AVRO, ...)
- Complex Data Types (Maps, Structs, Arrays)
- Schema on Read
- Python, R, Jupyter

**WEAKNESSES**
- Complexity
- Duplicated effort
- Division of BI and Data Science

© 2020 All rights reserved
Manage Data Life Cycle with Data Formats

LOW LATENCY
STREAMING DATA
Application data
- Web clicks
- Logs
- Sensors
- Operational metrics
- User tracking
- Geo-location

BATCH
CONTEXTUAL DATA
Files
- Weather
- Geo

TRANSACTIONAL DATA
Application Data
- OLTP/ODS

STREAM PROCESSING
Raw Data
Many Formats
Ingestion, ELT, Data Prep
ELT, Data Prep
Hot data, fast data ROS
Import, Export, Join
Cold data, historical data
Parquet

ON-PREMISES, HYBRID, CLOUD OR MULTI-CLOUD

SQL

BUSINESS INTELLIGENCE + DATA SCIENCE

© 2020 All rights reserved
Separate Compute from Storage

**LOW LATENCY**
STREAMING DATA
- Application data
  - Web clicks
  - Logs
  - Sensors
  - Operational metrics
  - User tracking
  - Geo-location

**BATCH**
CONTEXTUAL DATA
- Files
- Weather
- Geo

TRANSACTIONAL DATA
- Application Data
- OLTP/ODS

**STREAM PROCESSING**
- Batch ETL
- Fast ELT
- Ingestion, ELT, Data Prep

**Data Formats**
- Parquet
- ORC
- ROS
- JSON

**APPLICATION DATA**
- Batch ETL
- Fast ELT

**CONTEXTUAL DATA**
- Application Data
- OLTP/ODS

**LOW LATENCY**
- Application data
  - Web clicks
  - Logs
  - Sensors
  - Operational metrics
  - User tracking
  - Geo-location

**BATCH**
- Contextual data
  - Files
  - Weather
  - Geo

**TRANSACTIONAL DATA**
- Application Data
- OLTP/ODS

**STREAM PROCESSING**
- Batch ETL
- Fast ELT
- Ingestion, ELT, Data Prep

**DATA FORMATS**
- Parquet
- ORC
- ROS
- JSON

**APPLICATION DATA**
- Batch ETL
- Fast ELT

**CONTEXTUAL DATA**
- Application Data
- OLTP/ODS

**LOW LATENCY**
- Application data
  - Web clicks
  - Logs
  - Sensors
  - Operational metrics
  - User tracking
  - Geo-location

**BATCH**
- Contextual data
  - Files
  - Weather
  - Geo

**TRANSACTIONAL DATA**
- Application Data
- OLTP/ODS

**STREAM PROCESSING**
- Batch ETL
- Fast ELT
- Ingestion, ELT, Data Prep

**DATA FORMATS**
- Parquet
- ORC
- ROS
- JSON

**APPLICATION DATA**
- Batch ETL
- Fast ELT

**CONTEXTUAL DATA**
- Application Data
- OLTP/ODS

**LOW LATENCY**
- Application data
  - Web clicks
  - Logs
  - Sensors
  - Operational metrics
  - User tracking
  - Geo-location

**BATCH**
- Contextual data
  - Files
  - Weather
  - Geo

**TRANSACTIONAL DATA**
- Application Data
- OLTP/ODS

**STREAM PROCESSING**
- Batch ETL
- Fast ELT
- Ingestion, ELT, Data Prep

**DATA FORMATS**
- Parquet
- ORC
- ROS
- JSON

**APPLICATION DATA**
- Batch ETL
- Fast ELT

**CONTEXTUAL DATA**
- Application Data
- OLTP/ODS
Unify Analytics

LOW LATENCY
STREAMING DATA
Application data
- Web clicks
- Logs
- Sensors
- Operational metrics
- User tracking
- Geo-location

STREAM PROCESSING

Data Formats
- Kafka
- Amazon Redshift
- Parquet
- ORC
- ROS
- JSON

Batch ETL

Batch ET, ELT

Contextual Data

BATCH

CON Textual DATA
Files
- Weather
- Geo

TRANSACTIONAL DATA
Application Data
- OLTP/ODS

SQL
- Python
- Jupyter
- R

Ingestion, ELT, Data Prep

ON-PREMISES, HYBRID, CLOUD OR MULTI-CLOUD

Business Intelligence + Data Science

Application data
- Web clicks
- Logs
- Sensors
- Operational metrics
- User tracking
- Geo-location

Batch ETL
- Files
- Weather
- Geo

Transactional Data
- Application Data
- OLTP/ODS

Ingestion, ELT, Data Prep

SQL
- Python
- Jupyter
- R

On-Premises, Hybrid, Cloud or Multi-Cloud

© 2020 All rights reserved
Isolate Workloads

LOW LATENCY
STREAMING DATA
- Application data
  - Web clicks
  - Logs
  - Sensors
  - Operational metrics
  - User tracking
  - Geo-location

STREAM PROCESSING
- kafka
- Amazon Redshift
- Parquet
- ROS
- {JSON}

Isolated Compute Workloads
- Ingestion, ELT, Data Prep
- Machine Learning
- Reporting
- Ad Hoc Queries

SQL
- python
- jupyter
- R

BATCH
CONTEXTUAL DATA
- Files
- Weather
- Geo

TRANSACTIONAL DATA
APPLICATION DATA
- OLTP/ODS

ON-PREMISES, HYBRID, CLOUD OR MULTI-CLOUD

APPLICATIONS
- Hadoop
- HDFS
- =

© 2020 All rights reserved
Unified Analytics Platform

**LOW LATENCY**
- **STREAMING DATA**
  - Application data
  - Web clicks
  - Logs
  - Sensors
  - Operational metrics
  - User tracking
  - Geo-location

**BATCH**
- **CONTEXTUAL DATA**
  - Files
  - Weather
  - Geo

**TRANSACTIONAL DATA**
- Application Data
- OLTP/ODS

**DATA FORMATS**
- Parquet
- ORC
- ROS
- JSON

**STREAM PROCESSING**
- kafka
- amazon

**ISOLATED COMPUTE WORKLOADS**
- Ingestion, ELT, Data Prep
- Machine Learning
- Reporting
- Ad Hoc Queries
- Model Evaluation, Deployment, Management

**ON-PREMISES, HYBRID, CLOUD OR MULTI-CLOUD**

**BUSINESS INTELLIGENCE + DATA SCIENCE**
- SQL
- Python
- R

**INGESTION, ELT, DATA PREP**
- Batch ETL
- Fast ELT

**CONTEXTUAL DATA**
- Batch ETL
- Fast ELT

**TRANSACTIONAL DATA**
- Files
- Weather
- Geo
Unified Analytics Warehouse Strengths

**DATA WAREHOUSE**
- Reporting/Business Intelligence
- High Performance
- High Concurrency
- Reliability
- Security
- Governance

**Unified Analytics**
- SQL
- BI Visualization Tools
- Python
- Jupyter
- R

**DATA LAKE**
- Machine Learning/Data Science
- Unlimited Scale
- Streaming Data
- Semi-Structured Data (JSON, AVRO, …)
- Complex Data Types (Maps, Structs, Arrays)
- Schema on Read
**EXTREMELY LOW LATENCY**
- Real-time bidding data

**LOW LATENCY**
- Third-party data

**BATCH**
- Contextual data

---

Unified Analytics Platform

**Primary Sub-Cluster**
- Data Ingest
- ELT / Aggregation

**Ephemeral Sub-Clusters**
- Reporting
- Reporting
- Ad Hoc BI Queries
- Machine Learning

---

Hot Fast Data

Big Stored Data

---

**SQL**

**APPLICATION**
- ML APPLIED

**BUSINESS INTELLIGENCE**
- ML TRAINING
- REPORTING

---

**AWS**
EMA Radar Report: Unified Analytics Warehouse
A Guide for Investing in Unified Analytics

Request your copy today:
Thank you

Paige Roberts
Open Source Relations Manager
E: Paige.Roberts@vertica.com