VERTICA

Cloud, On-Prem, and People

A Practical Guide to Advanced Analytics and ML Adoption

Bryan Whitmore, Chief Field Technologist, Vertica

MICRO FOCUS

By way of Introduction

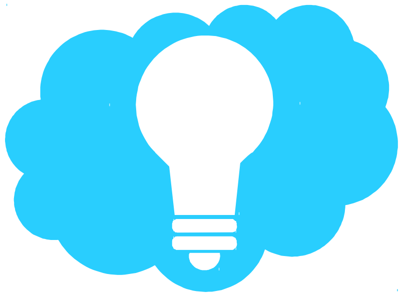
- **Vertica is 11+ years and growing**
  - I've been with Vertica 7+ years
- **We sit between data sources, orchestration tools, and consumers**
  - Storing, Organizing, Transforming, and providing Offload
  - Some of the biggest, but also many small, including OEM/Embeded
  - Many User Stories to share
- **Varying Objectives and Maturity**
  - But everybody wants to leverage what they have and gain flexibility

4 Common Needs

Data Discovery Projects

Understand value before committing big $$$

Unlocking “Dark” Data



Reduce Barriers to Leveraging Underutilized Data

Performance at Scale

Cost Effective, Elastic, Capable

Overcoming Silos

Eliminate Redundant Pipelines & Copies

Approaches to Discovery

- **Connect my Laptop to Data, Produce a Workbook**
  - Low barrier to entry, potential versioning issues – a workbook is not a sandbox

Approaches to Discovery

- **Connect my Laptop to Data, Produce a Workbook**
  - Low barrier to entry, potential versioning issues – a workbook is not a sandbox
- **Self-Service BI connected to a Database**
  - Great for democratizing data across lines of business
  - Leverage unique points of view
  - Later phase, since the value is established and the data modeled

Approaches to Discovery

- **Connect my Laptop to Data, Produce a Workbook**
  - Low barrier to entry, potential versioning issues – a workbook is not a sandbox
- **Self-Service BI connected to a Database**
  - Great for democratizing data across lines of business
  - Leverage unique points of view
  - Later phase, since the value is established and the data modeled
- **Connect an Analysis Tool to a Data Lake (HDFS, S3, etc)**
  - Prices start as low as 50 cents per hour

Approaches to Discovery

- **Jupiter, R Studio, Spark, SAS Integration Story**
  - Getting the most from skilled people - empower their tools of choice
  - Data Gravity: Infrastructure to offload laptops, pushing it down to data repo

Approaches to Discovery

- **Jupiter, R Studio, Spark, SAS Integration Story**
  - Getting the most from skilled people - empower their tools of choice
  - Data Gravity: Infrastructure to offload laptops, pushing it down to data repo
- **Get a Handle on Versioning**
  - Workbooks checked in, not emailed around

Approaches to Discovery

- **Jupiter, R Studio, Spark, SAS Integration Story**
  - Getting the most from skilled people - empower their tools of choice
  - Data Gravity: Infrastructure to offload laptops, pushing it down to data repo
- **Get a Handle on Versioning**
  - Workbooks checked in, not emailed around
- **Self-Service means knowing the answer during a meeting**
  - Why follow up when you can do it live
  - Individual Discovery becomes Company Culture

Unlocking Dark Data

- **Turns out, SQL was what we wanted the whole time** (most of the time)
  - Everything from Excel, to emerging Visualization and Modeling tools, to in-house applications, can talk to a database
  - Query Engine access to Data Lakes is in high demand

Unlocking Dark Data

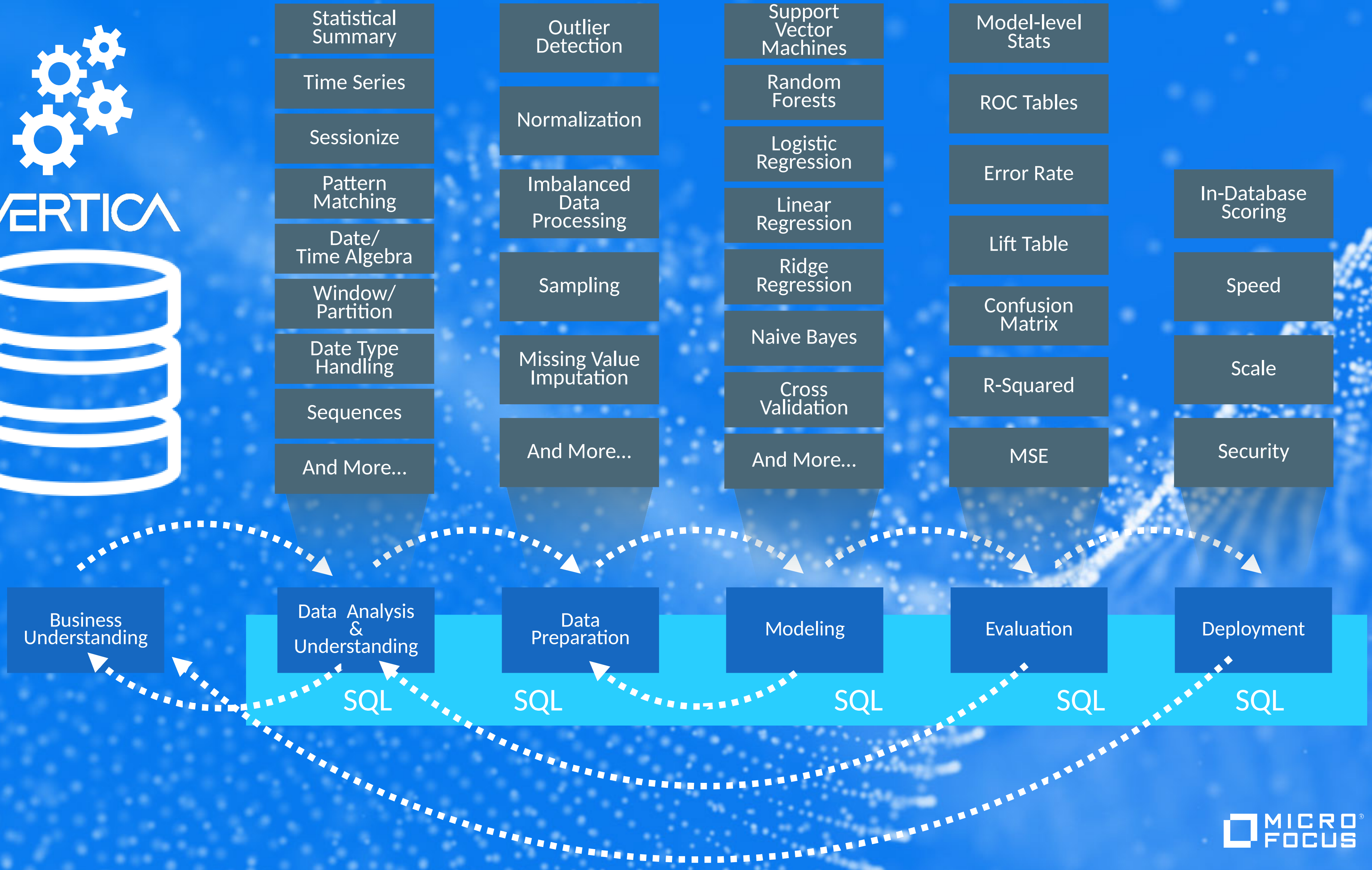
- **Turns out, SQL was what we wanted the whole time** (most of the time)
  - Everything from Excel, to emerging Visualization and Modeling tools, to in-house applications, can talk to a database
  - Query Engine access to Data Lakes is in high demand
- **Sometimes Latency Matters**
  - What are customers/suppliers experiencing today, within the last hour, or 5 minutes?
  - **Streaming** (eg Kafka, Avro, JSON) or **Files** on lake / object store (flat, JSON, ORC, Parquet)

Unlocking Dark Data

- **Turns out, SQL was what we wanted the whole time** (most of the time)
  - Everything from Excel, to emerging Visualization and Modeling tools, to in-house applications, can talk to a database
  - Query Engine access to Data Lakes is in high demand
- **Sometimes Latency Matters**
  - What are customers/suppliers experiencing today, within the last hour, or 5 minutes?
  - **Streaming** (eg Kafka, Avro, JSON) or **Files** on lake / object store (flat, JSON, ORC, Parquet)
- **Spark, Python are Hot**
  - Practices are becoming established and people are growing their skills

Vertica Machine Learning Process Flow

- Machine Learning
- Speed
- ANSI SQL
- Scalability
- Massively Parallel Processing
- Deploy Anywhere



Business Understanding

Data Analysis & Understanding (SQL)
- Statistical Summary
- Time Series
- Sessionize
- Pattern Matching
- Date/ Time Algebra
- Window/ Partition
- Date Type Handling
- Sequences
- And More…

Data Preparation (SQL)
- Outlier Detection
- Normalization
- Imbalanced Data Processing
- Sampling
- Missing Value Imputation
- And More…

Modeling (SQL)
- Support Vector Machines
- Random Forests
- Logistic Regression
- Linear Regression
- Ridge Regression
- Naive Bayes
- Cross Validation
- And More…

Evaluation (SQL)
- Model-level Stats
- ROC Tables
- Error Rate
- Lift Table
- Confusion Matrix
- R-Squared
- MSE

Deployment (SQL)
- In-Database Scoring
- Speed
- Scale
- Security

Performance at Scale

- **Sometimes you can just throw Money at a Problem**
  - If a platform does the job at 10x, 100x, 1000x scale, even if it isn't efficient, so long as it is linear
  - The economics may work, especially if it avoids re-implementing (People cost more than computers)
- **Other times Efficiency Matters**
  - Internet-Scale companies often code entire infrastructures to operate cost-effectively
  - Some of those tools are Open Source
- **Elastic Consumption is a Given, even On-Prem**
  - There are cost savings to buying in bulk, but purchasing ahead of demand is becoming the exception rather than the norm

Performance at Scale

- **Sometimes you can just throw Money at a Problem**
  - If a platform does the job at 10x, 100x, 1000x scale, even if it isn't efficient, so long as it is linear
  - The economics may work, especially if it avoids re-implementing (People cost more than computers)

Performance at Scale

- **Sometimes you can just throw Money at a Problem**
  - If a platform does the job at 10x, 100x, 1000x scale, even if it isn't efficient, so long as it is linear
  - The economics may work, especially if it avoids re-implementing (People cost more than computers)
- **Other times Efficiency Matters**
  - Internet-Scale companies often code entire infrastructures to operate cost-effectively
  - Some of those tools are Open Source

Overcoming Silos

- **Every Line of Business wants the Purchasing Data**
  - …and the Browsing, and the industry comps, and the long-tail Seasonal
  - But they each have their own derived metrics, ETL pipelines, and SLA

Overcoming Silos

- **Every Line of Business wants the Purchasing Data**
  - …and the Browsing, and the industry comps, and the long-tail Seasonal
  - But they each have their own derived metrics, ETL pipelines, and SLA
- **All the Reference Data in One Place**
  - EDW promised that, but required so much management nobody could touch it
  - Data Lakes + Query Engines consolidate Data Marts, but batch didn’t cover SLA

Overcoming Silos

- **Every Line of Business wants the Purchasing Data**
  - …and the Browsing, and the industry comps, and the long-tail Seasonal
  - But they each have their own derived metrics, ETL pipelines, and SLA
- **All the Reference Data in One Place**
  - EDW promised that, but required so much management nobody could touch it
  - Data Lakes + Query Engines consolidate Data Marts, but batch didn’t cover SLA
- **Cloud-Native Databases can, but can get expensive**
  - $10,000 for a query is what one customer shared with me …we can do better
  - Our Vision: highly leveraged tables in-database, everything else as external tables in cheap object storage

Overcoming Silos

- **What Security Features do you require?**
  - Multi-Realm Kerberos, Encryption, Multi-Factor Authentication
  - Access Policy and Governance integration

Overcoming Silos

- **What Security Features do you require?**
  - Multi-Realm Kerberos, Encryption, Multi-Factor Authentication
  - Access Policy and Governance integration
- **Can Tuning improve economics / SLA – can developers do it?**
  - Design automation, Schema evolution, Profiling, **iterative during development**
  - User Defined Functions, including ML, as calculated column values
  - Workload Isolation (isn’t a given just because cloud)

Overcoming Silos

- **What Security Features do you require?**
  - Multi-Realm Kerberos, Encryption, Multi-Factor Authentication
  - Access Policy and Governance integration
- **Can Tuning improve economics / SLA – can developers do it?**
  - Design automation, Schema evolution, Profiling, **iterative during development**
  - User Defined Functions, including ML, as calculated column values
  - Workload Isolation (isn’t a given just because cloud)
- **Dependency Graphs and Access Counters**
  - Who depends on data produced by Group A, and how leveraged is that data

Thank You