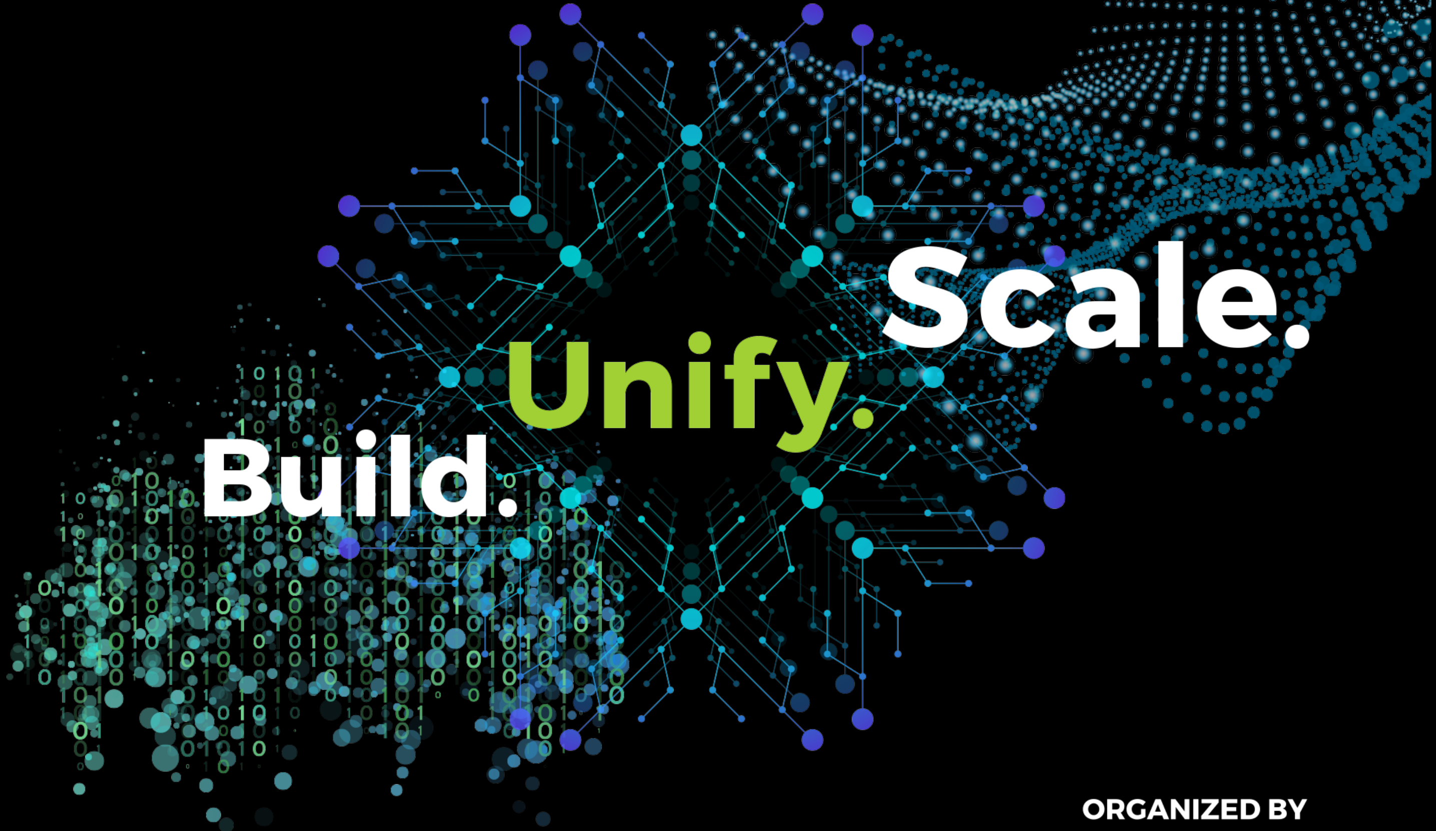
SPARK+AI
SUMMIT 2019



Build. Unify. Scale.

WIFI SSID:SparkAISummit | Password: UnifiedAnalytics

ORGANIZED BY
databricks

SPARK+AI
SUMMIT 2019

Big Data Journey to Create the 360 View of the Consumer : *Data Driven Strategies with Data Lake and Databricks*

Jyoti P. Mohapatra, Altria

Ramesh Ketha, Capgemini

#UnifiedAnalytics #SparkAISummit

About Altria

Altria's companies have a strong American heritage stretching back more than 180 years.

Altria Group holds diversified positions across tobacco, alcohol and cannabis. Through our wholly-owned subsidiaries and strategic investments in other companies, we seek to provide category-leading choices to adult consumers, while returning maximum value to shareholders through dividends and growth.

We are a FORTUNE 200 company, proud to call Richmond, Virginia our home. Our people and companies address tough industry issues, like reducing the health effects of tobacco use and preventing underage tobacco use. And we focus on strengthening the communities where we live and work.

Our Mission & Values

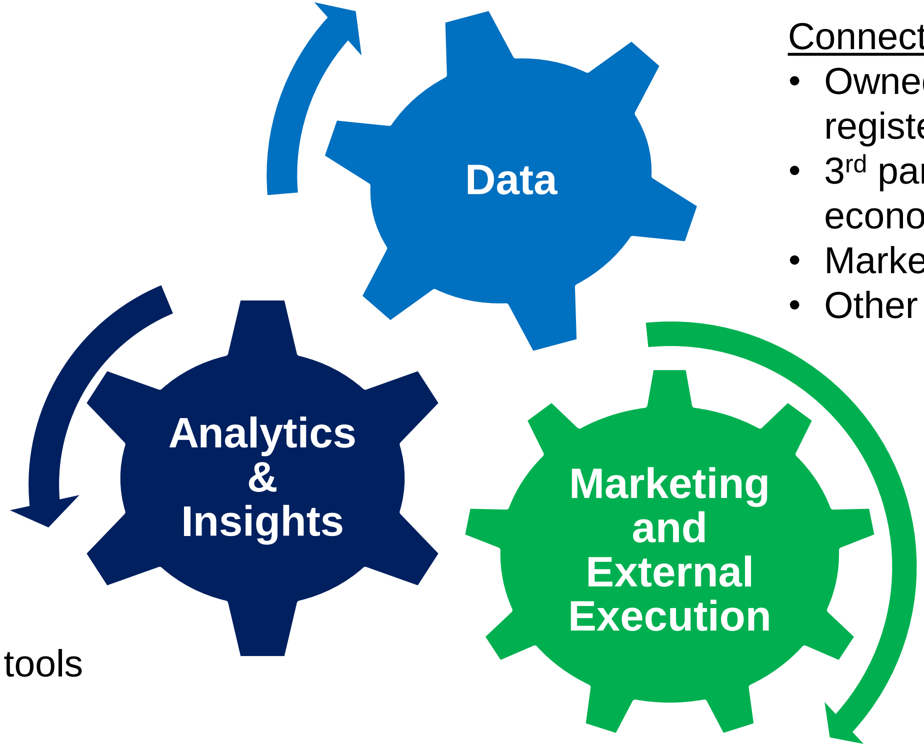
Is to own and develop financially disciplined businesses that are leaders in responsibly providing adult tobacco and wine consumers with superior branded products.

Our Values guide our behavior as we pursue our Mission and our business strategies : Integrity, Trust and Respect Passion to Succeed, Executing with Quality, Driving Creativity into Everything We Do, Sharing with Others.

Context

- Data is a competitive advantage for Altria
  - Adult Consumer Database
  - Marketplace Information
  - Trade Program
- Access to and use of new adult consumer information and sources of data are increasing
- Very competitive and regulated market
- Growth impact seen at companies that inject analytics into their operations
- Building up and connecting data will drive better insights and continued advantage for Altria

Mission



Connected Data

- Owned ATC 21+ data who are age verified, registered and Opt-In
- 3rd party data (e.g. public data: census, economic data etc.)
- Marketplace POS Scan data
- Other Altria operational data

Analytics Roadmap

- ATC Understanding
- Precise Value and Equity Delivery
- Enable Salesforce
- Product Innovation and Regulatory Approval
- External Engagement

Synthesis

- Analytical tools
- People
- Process

Business Initiatives

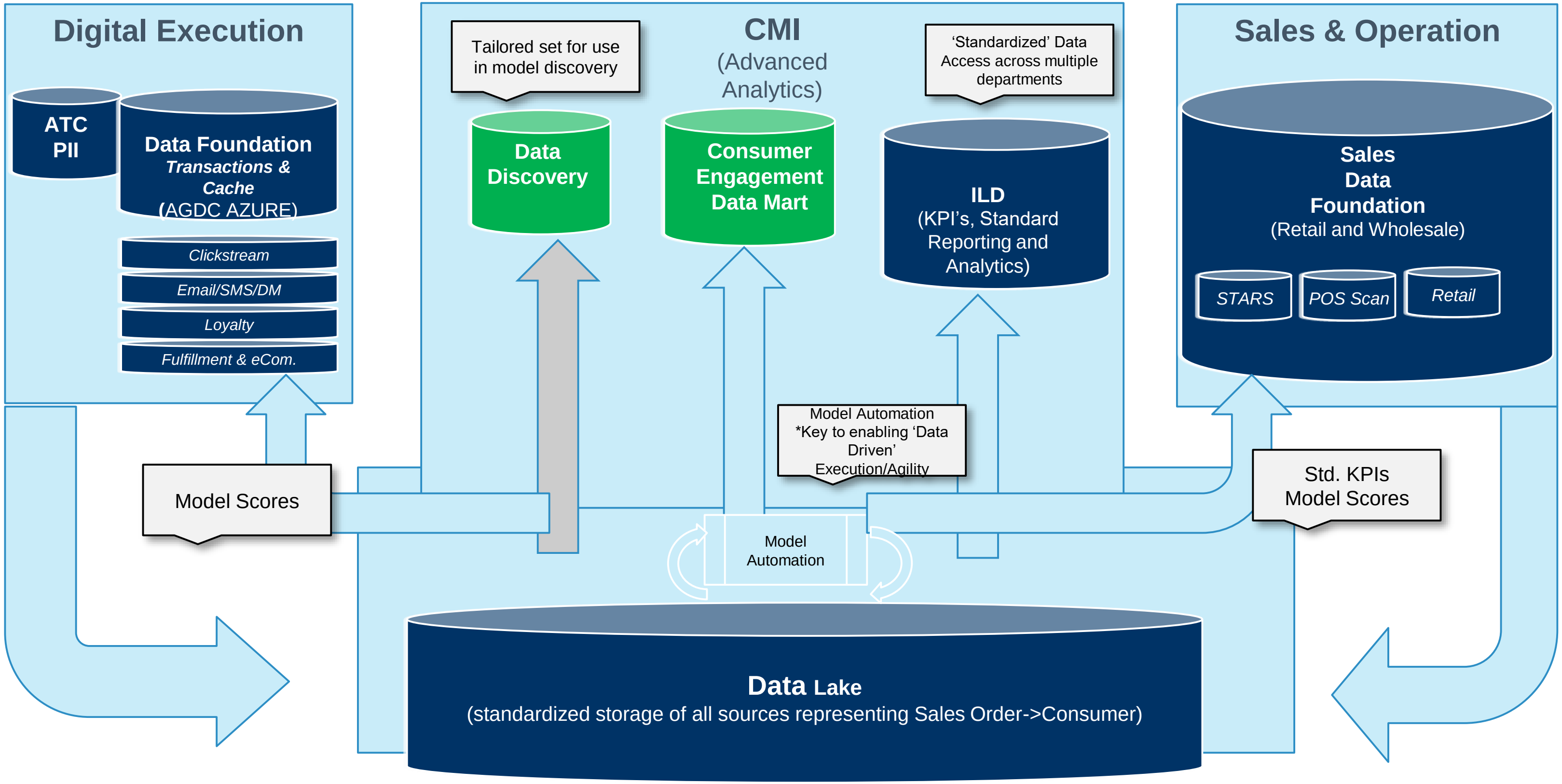
- Digital Transformation
  - Adult Consumer 360 and Personalization
  - Marlboro Rewards Launch
  - Market Basket Analysis
  - Precise Value and Equity Delivery
  - Product Innovation and Regulatory Approval

- Data Velocity and Volume
  - Growth in POS Scan Data
  - Trade Program Management
  - Competitive Products
  - Trade Payments
- Sales Application Cloud Migration
  - Reduce the Data footprint On-Premise
  - Data Interfaces ,Pipelines and Process rebuild
  - Applications Transactional sync
- Data Governance and Stewardship and Unified Access

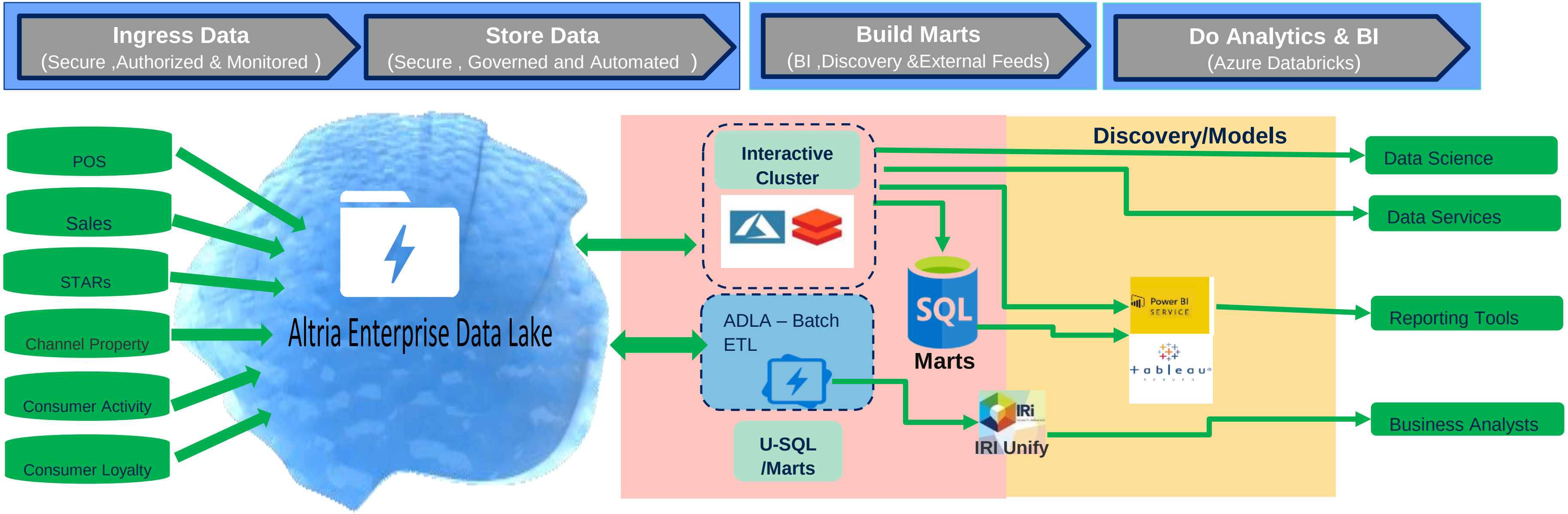
Challenges

- Data Content stored in disparate sources
- Limited integrated view of adult consumers and cross channel activities
- Cumbersome, slow data access
- Asynchronous data exchange with suppliers and adult consumer touchpoints (e.g., Email, SMS)
- Limited analytics capabilities, e.g., real-time personalization, coupon optimization, cross channel harmonization, experimentation
- Siloed architecture that limits cross-channel experiences and scalability

Data Landscape



Enterprise Data Lake Journey



Ingress Data
(Secure ,Authorized & Monitored)

Store Data
(Secure , Governed and Automated)

Build Marts
(BI ,Discovery &External Feeds)

Do Analytics & BI
(Azure Databricks)

Data Lake Services

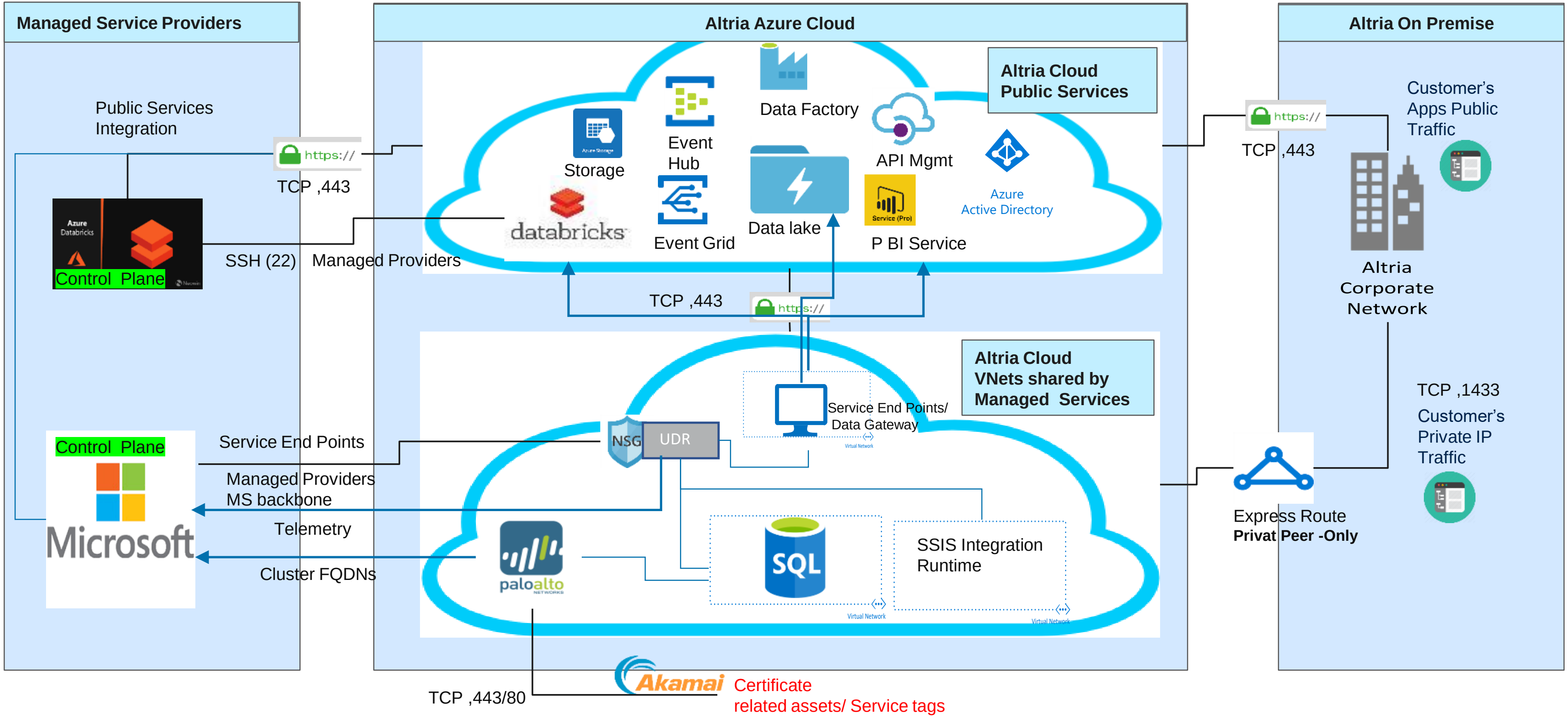
- *Data Enrollment Program (**DEP**) – Automated new data providers, Notifications and Monitoring.*
- *Data Analytics Services (**DAS**) – Data Marts , Modeled , Visualization*
- *Data Governance Service (**DGS**) – Audit Ingress , Stage and Egress*
- *Infrastructure Data Service (**IDS**) – Archive/Stage Only*
- *Pre-modeled Data Host Services (**PDHS**) – Stage Pre-modeled/Aggregated data for Visualization*

Altria Design Principles

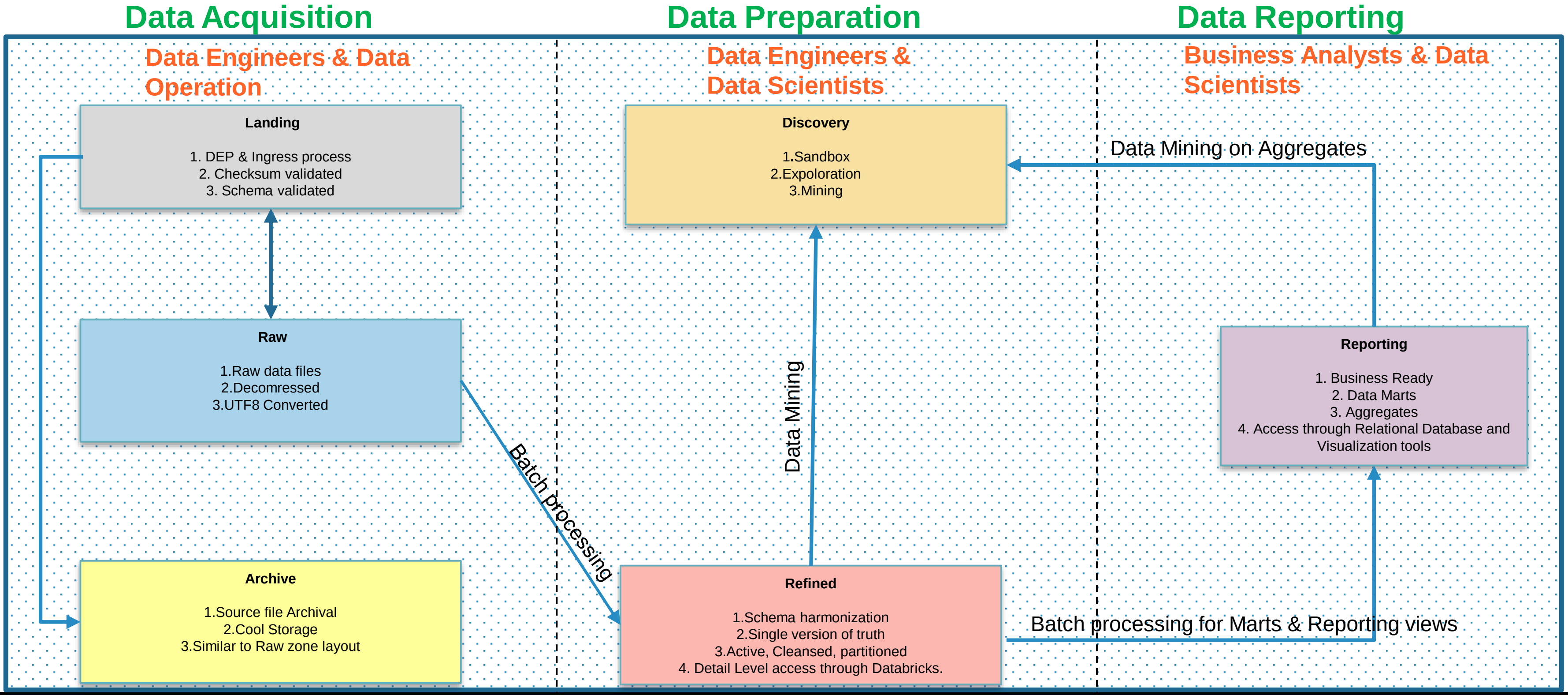
- PaaS first Solution
- Security & Governance inline with Enterprise Architectural guidelines
- Secured Azure Cloud Environment, ‘Private Peer’ Express route only from On-Premise to Cloud.
- PAAS Service provider Must have
  - Identity Management (AAD)
  - Approved Networking & Security
  - Vulnerability & Audit reports

- Consolidated datasets in central location (without Personally Identifiable Information)
- Data resides within Altria Subscription
- Enable a ‘Single Source of Truth’
- Enable analytics
  - Quick and easy access to information
  - Leverage power of cloud computing to enable machine learning / advanced analytics
- Governed by Information and Insights Initiative Data Principles

Azure PaaS Reference Architecture



Data Lake Data Flow Strategy– Multi Zone Implementation



Data Acquisition

Data Engineers & Data Operation

Landing

1. DEP & Ingress process
2. Checksum validated
3. Schema validated

Raw

1.Raw data files
2.Decomressed
3.UTF8 Converted

Archive

1.Source file Archival
2.Cool Storage
3.Similar to Raw zone layout

Batch processing

Data Preparation

Data Engineers & Data Scientists

Discovery

1.Sandbox
2.Expoloration
3.Mining

Data Mining

Refined

1.Schema harmonization
2.Single version of truth
3.Active, Cleansed, partitioned
4. Detail Level access through Databricks.

Data Reporting

Business Analysts & Data Scientists

Data Mining on Aggregates

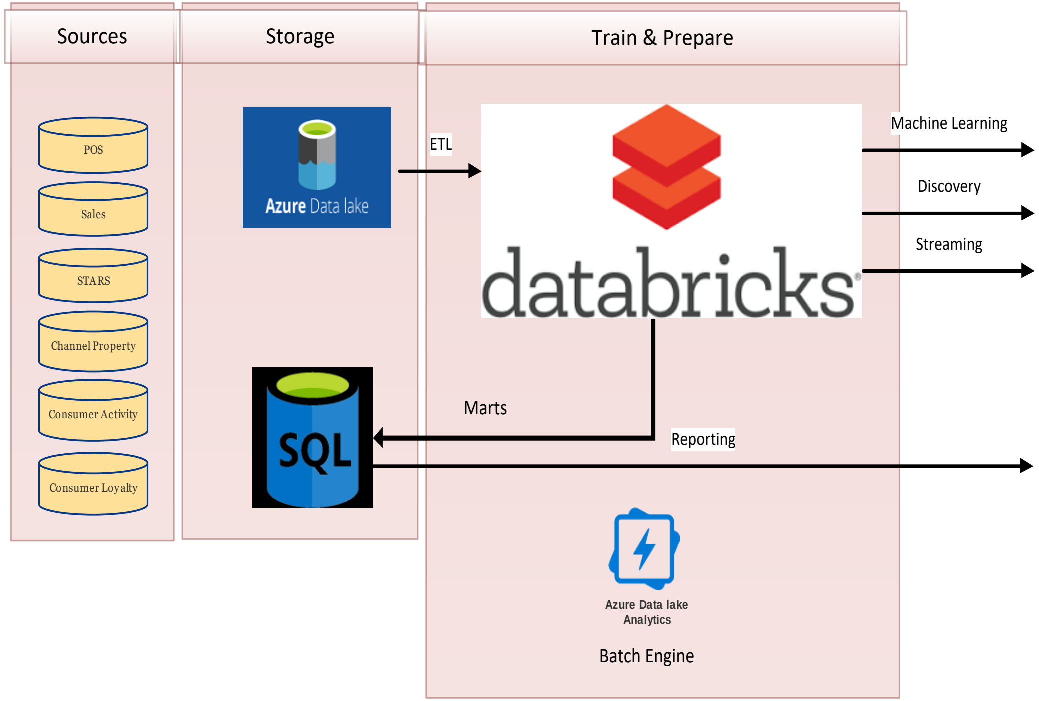
Reporting

1. Business Ready
2. Data Marts
3. Aggregates
4. Access through Relational Database and Visualization tools

Batch processing for Marts & Reporting views

Why Databricks on Azure

- Self-Service cluster management
- Easy to configure and all backend services Managed by Databricks
- Integration with Azure Identity Management
- Easy integration with Azure Data Lake, Storage and SQL Database and other Azure native cloud services
- Major Contributor of Open Spark
- Excellent Support for Data Science Development languages like Python, R, Scala etc.
- Speed to market on Technology & Secured implementation for Hybrid access
- Collaborative Notebooks and Less Code Rewrites
- Full Suite of Data Transformation and ML capabilities including MLFlow



Implementation Challenges

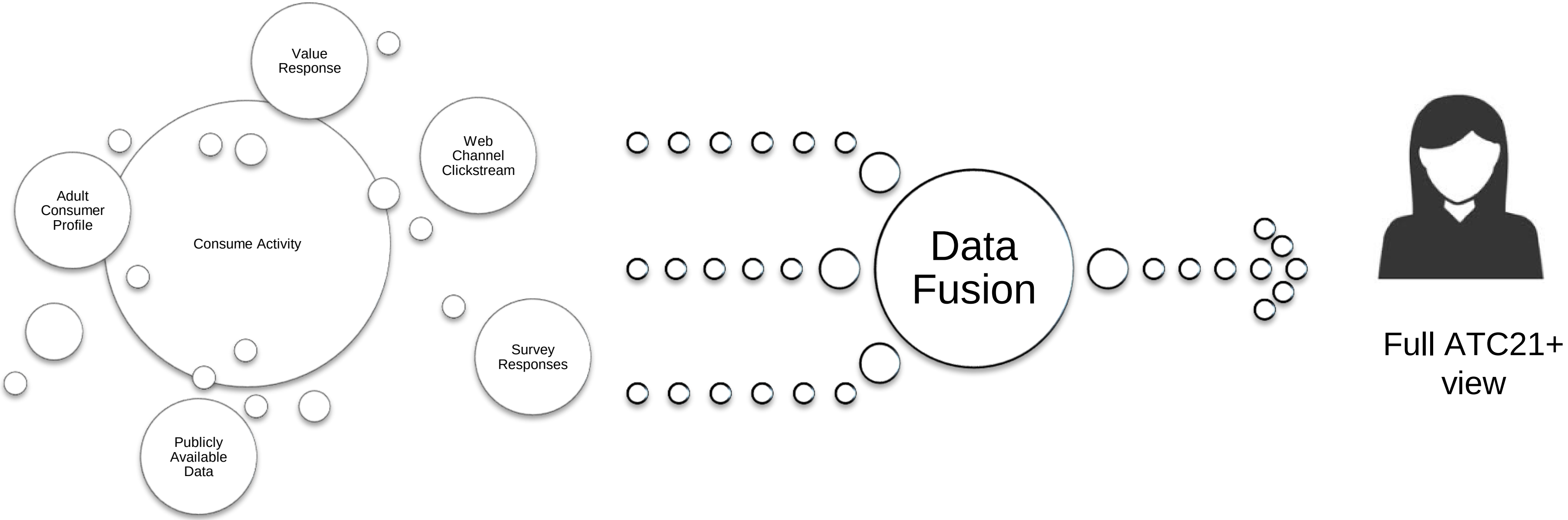
- Solution Involve Hybrid PaaS offerings
- New era of Altria Cloud VNETs being shared by Service Providers for managed services
- Routing trust for Managed Services without Firewall Appliances
- Hosting Public IP's in Altria cloud VNETs and no Express route
- Multiple Key Stake holders involvement for Security firewall and Networking landscape
- Altria Networking Landscape is evolving - So many moving parts
- Subject Matter Experts new to Azure
- New Tools being matured (Single Sign-on, Security and Networking, Evolving Azure storage Gen2)
- Legacy SQL Users transition to Notebooks, new skills working with cloud tools

Success Measure

- Data Lake which includes structured and unstructured data to create a consumer 360
- Variety of data storage types to pre-process data from the Data Lake to support faster and efficient data access
- Synchronous data exchange with suppliers, retailers, and digital properties via Restful APIs
- Robust unified analytics platform using Databricks to support new capabilities, e.g., advanced analytics , optimization, and experimentation
- Single data repository and engine that has enormous processing power and ability to handle concurrent tasks

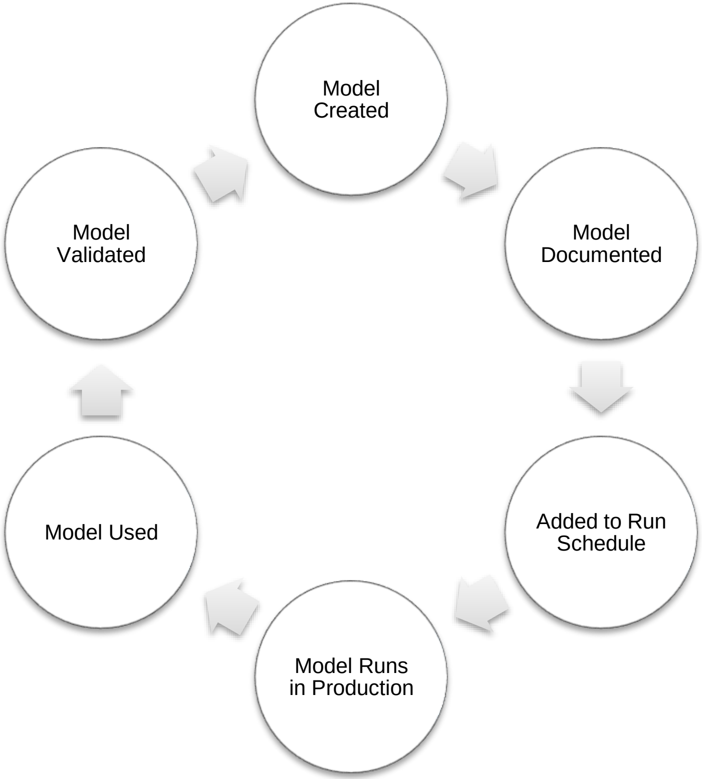
Success Measure – Model Scoring (Data Fusion)

Aggregates and consolidates, to a single view, all of the *known* data about an ATC 21+ to model the *unknown* - Main component of Model Engine



Success Measure / Data Builder & Manager

- System integrated with Data Fusion to manage all aspects of the model life cycle
    - Model Documentation
    - Model Run Schedule
    - Model Scoring and Validation
- R , Spark SQL and PySpark interface to Data Fusion that greatly simplifies creating custom models
    - Generates 1,000+ time specific variables for modeling
    - Takes care of all data munging and cleaning
    - Creates first pass XGBoost and elastic net models
    - Standardizes variable creation for consistent use across Altria and within vendor network



Success Measure - Altria Model Engine

- In-house model management solution
- Rapidly build and install new models
- Technology agnostic (aligns with IS/Digital Machine Infrastructure)
- Formalized Model Documentation process
- QC all data in one place
- Flexibility to add new data as available

Learnings…

- Connectivity Issues can cause on-premise job failures & irreversible data since no atomic operations support Data Lake Gen1 - Build support failover and monitoring
- Data Lake folders case sensitive
- Data Lake Folder permissions inheritance with Service Principles
- ADLA (U-SQL) supported only UTF8 encoding . Evolved to support zipping and unzipping the files, schema validations but make sure it runs on single node.
- ADLA node limits , concurrent jobs and working with Parquet Compressed files.
- ADLA ,USQL doesn’t have inbuilt capabilities to extract files having different schema , Custom Extractors
- Files transfers move from Logic Apps to Data Factory . Logic Apps support up to 1 GB.

Learnings…

- Logic Apps Triggers inconsistence with X number of files & Size.
- Pipeline deployments with Power shell , make sure ADF2 & Power shell version on sync to deployments
- ADF V2, Databricks and Parameterized notebooks , Worked with MS & closed issues
- EventHub triggers can only handled through Function Apps and issues with long running jobs
- EventHub integration with only Azure blob and Data Lake and only AVRO.
- SQL Managed Instance no integration with Polybase & Data Lake
- Databricks , Driver Node (only 2GB) limitation with Pandas or Native R . Move to SparkR or SparklyR
- Databricks , Data Lake with Mount points and moved Pass through & session scope access points

We are reducing bureaucracy, decentralizing decision-making and more effectively using data analytics to drive strategy

SPARK+AI
SUMMIT 2019

DON’T FORGET TO RATE AND REVIEW THE SESSIONS

SEARCH SPARK + AI SUMMIT

Download on the App Store

GET IT ON Google Play