Verizon: Finance Data Lake Implementation as a Self Service Discovery Big Data Platform

Sreenath Akinepalli

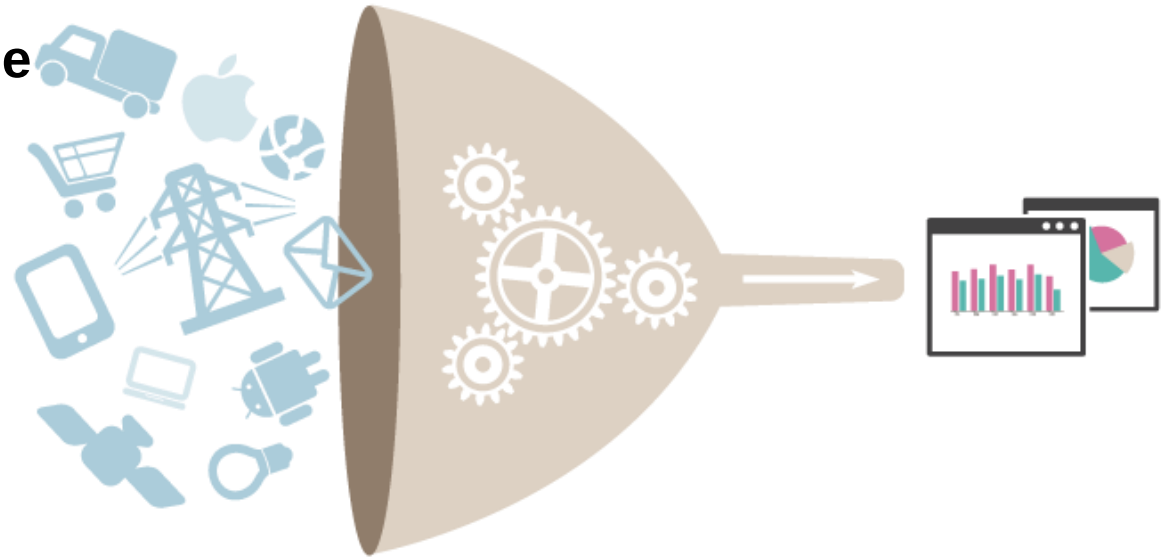
Sandeep Katuku

June 2017

verizon

Agenda

- Why Data Lake?
- Finance Data Lake Value
- Use Cases
- Architecture Overview
- Data Ingestion at Scale
- Data Validation
- Security
- Self Service Discovery
- Takeaways



verizon

Why Data Lake?

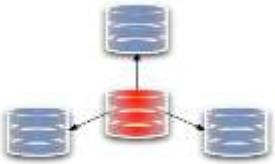
- **70% of time spent data gathering vs 30% analysis**
- **Data exists in multiple ERP systems and other silos**
- **Data Replication through point to point interfaces**
- **Lack of Normalization & Harmonization**
- **Data Latency**

verizon

Finance Data Lake Value

Centralized Enterprise Data Repository & Self-Service Discovery Platform

Simplifies access to raw ERP data



Eliminate data replication – lower TCO

Enable Data Share - reduce # of point to point integrations

Drive Data Archiving Strategy

Rationalize & harmonize master data

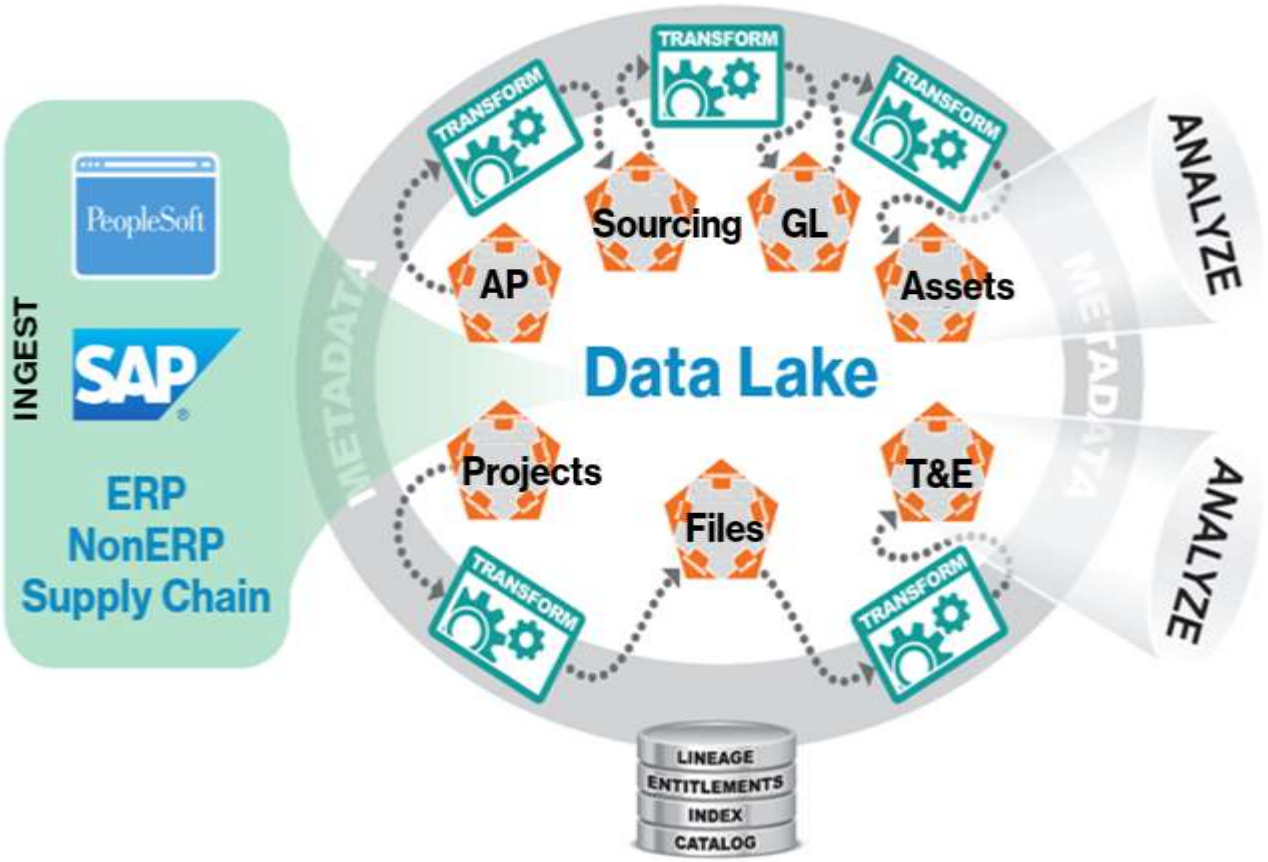
Centrally apply common business rules

Single set of Reporting & Analytical tools

Data Goverance & Security

verizon

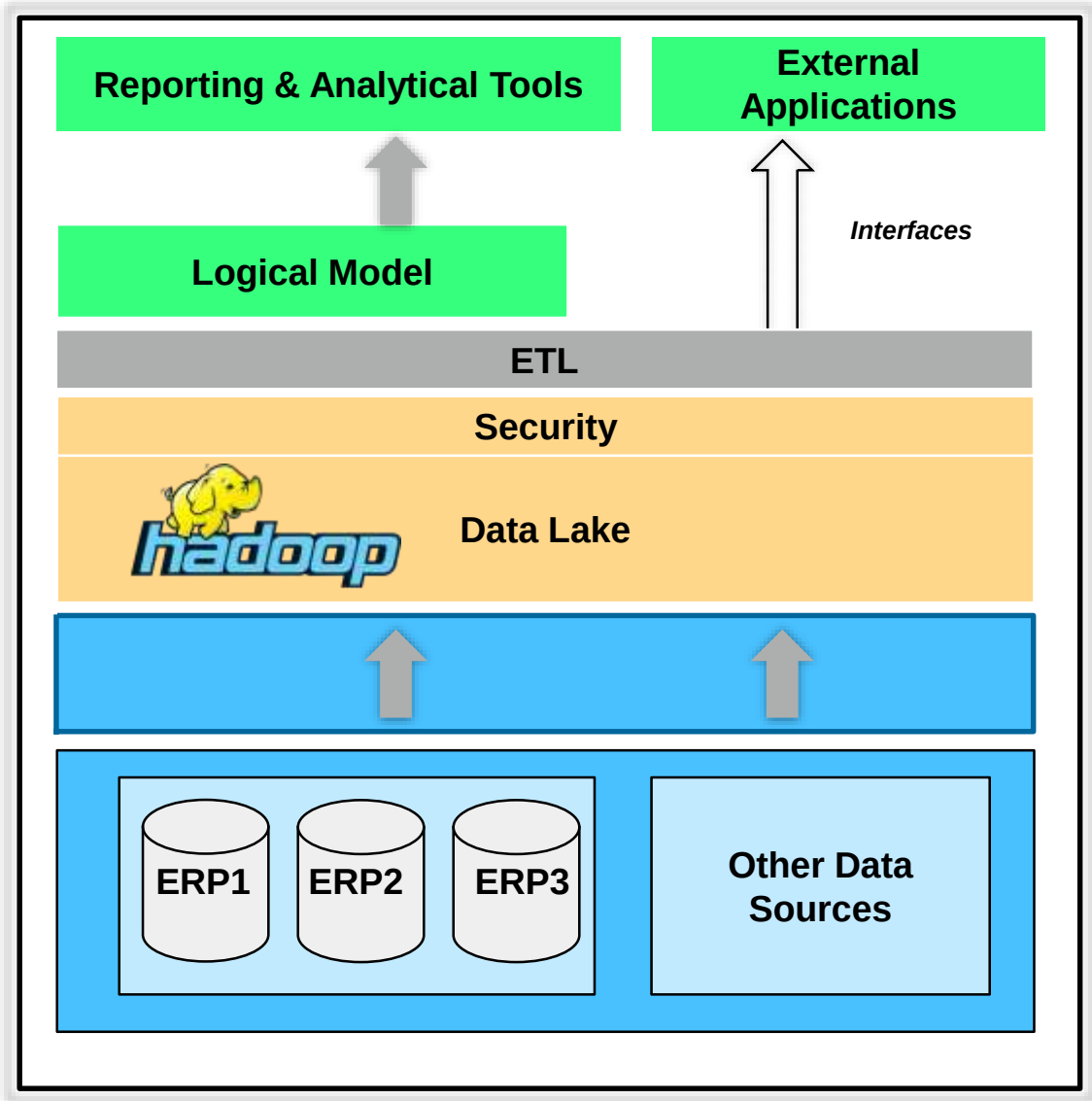
Use Cases



- **Accounts Payable (AP) Working Capital Analytics**
- **Historical DataMart**
- **Spend Analytics**
- **Labor Transformation**
- **Audit & Compliance**
- **Capital Reporting & Analytics**

verizon

Architecture Overview



Data Consumption

3

Semantic Model

Transformation

Governance

2

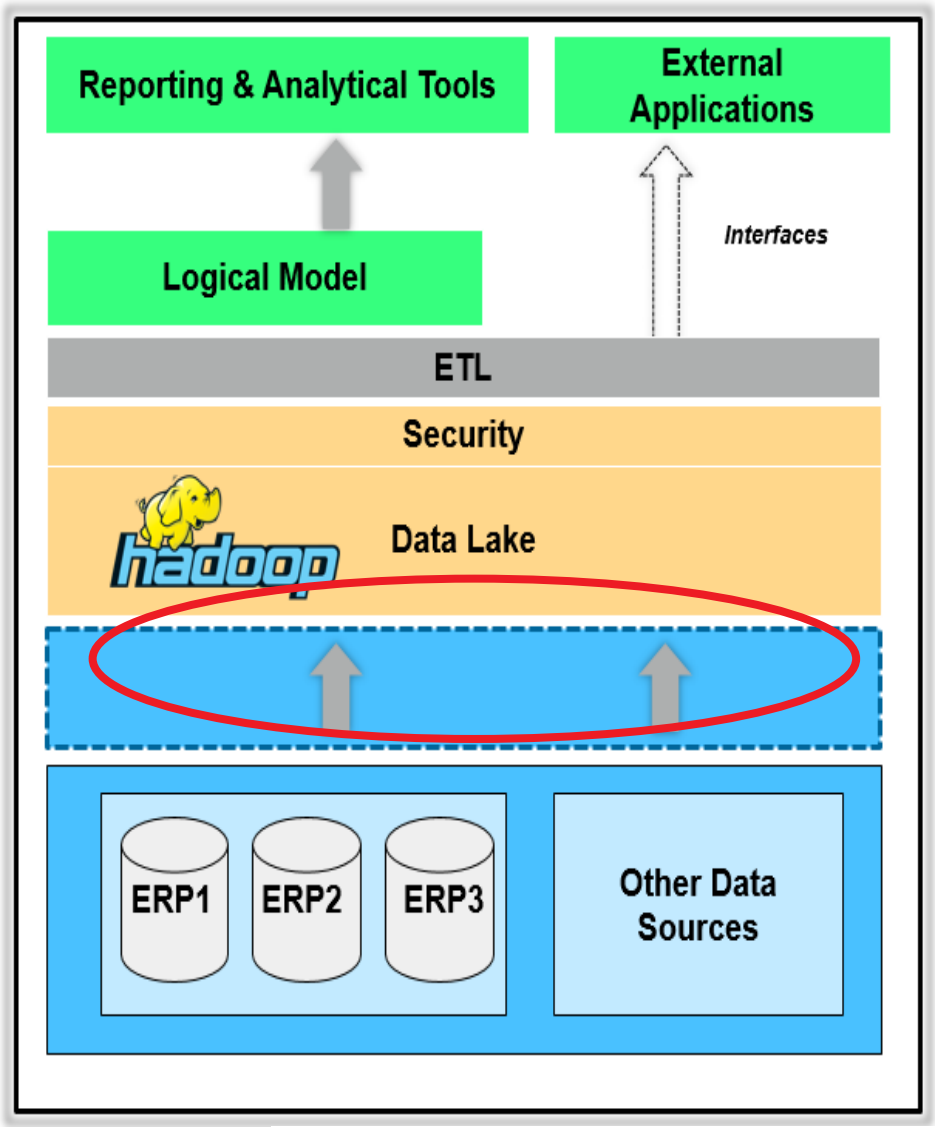
Store

Ingestion

1

Source Systems

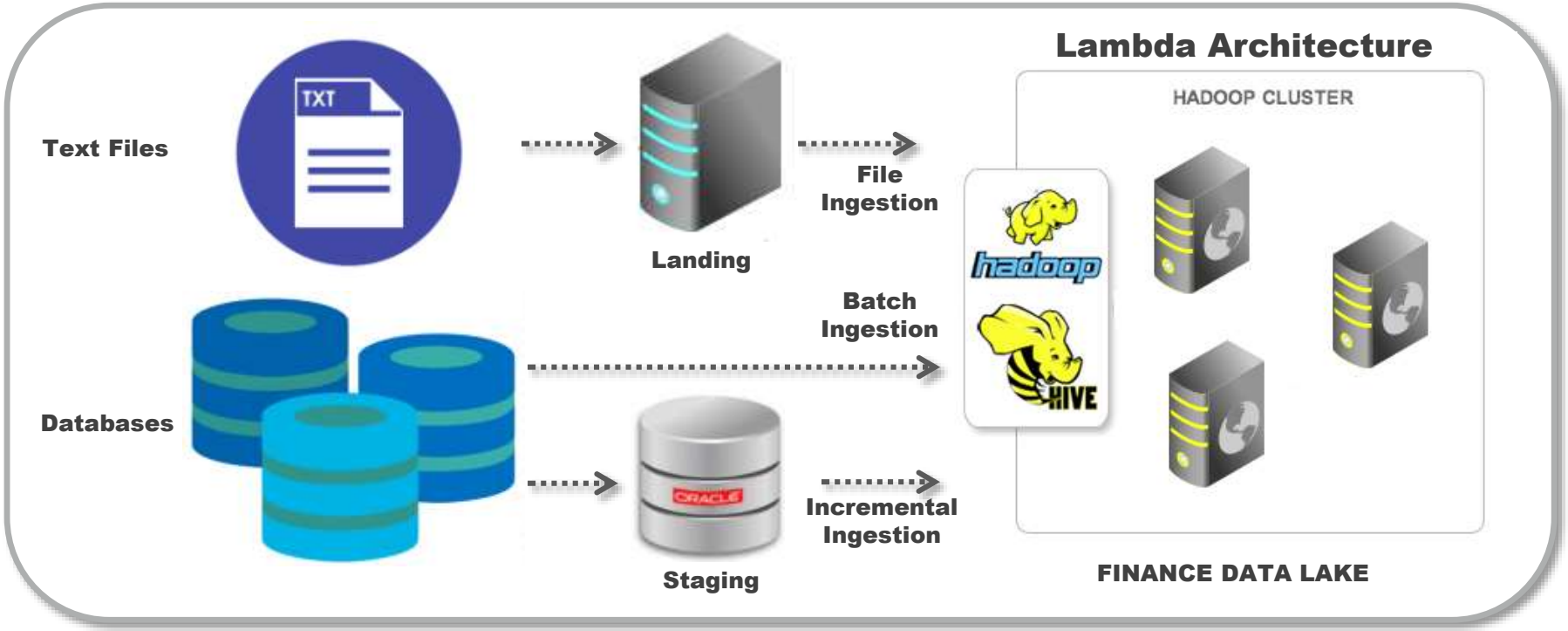
Data Ingestion - Design Success Factors



- Hadoop as Target
- Multiple Data Sources
- Transactional Source Systems (OLTP)
- ACID Limitations
- Different types of tables
- Ability to Scale to thousands of tables
- End to End Traceability
- Identifying the Tools

verizon

Data Ingestion at Scale



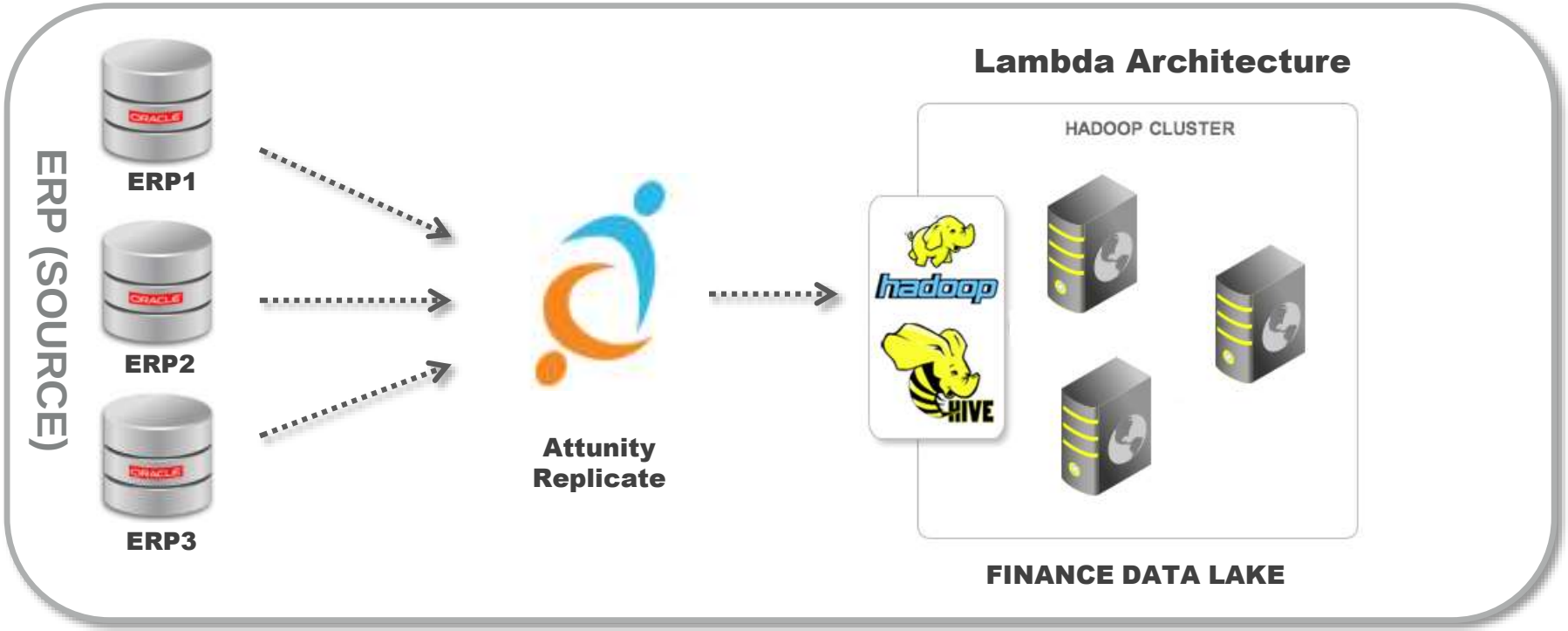
- Metadata Driven Design
- Dynamic Object Creation
- Supports File, Batch & Incremental Ingestion
- File & Batch Ingest Data directly to Hadoop
- Incremental data streams from Source to Staging
- Micro batch process moves data to Hadoop
- Data Merge using Lambda Architecture

verizon

Data Ingestion - Data Patterns

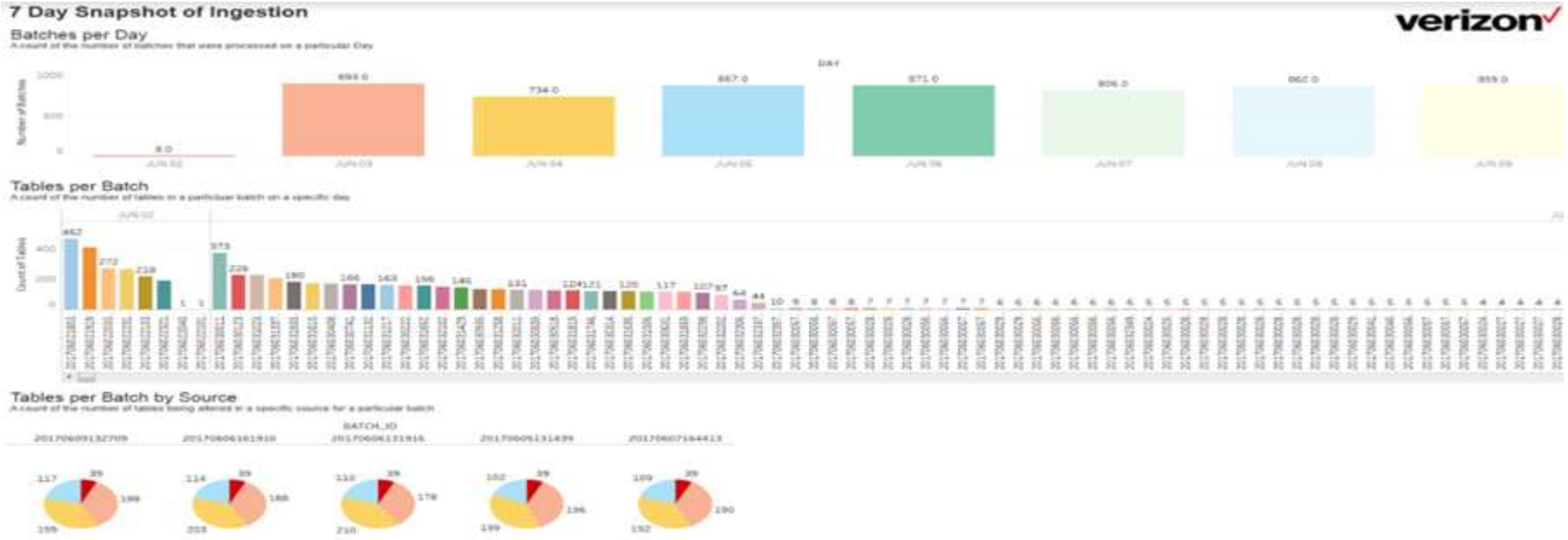
Challenge		Solution
Handling Deletes	→	Prior snapshots as deletes
Handling Updates	→	Updates as deletes and inserts
Primary Key updates	→	Prior snapshots as deletes
Concurrent Operations	→	Using transaction id
Batch Operations	→	Configuration to capture truncates

Data Ingestion - Enhancement

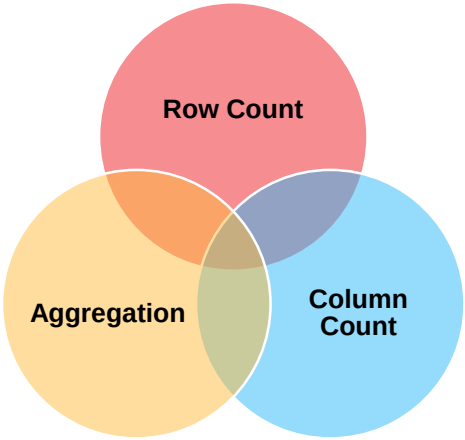


- Simplified Architecture
- Ingest Data directly to Hadoop
- Supports all ERP tables
- Dynamic DDL changes from Source to Target

Data Validations



- 4000 Automated Validations
- Row & Column count Dashboards
- Source to Hadoop Comparisons
- Data Latency Dashboards
- Report Reconciliation



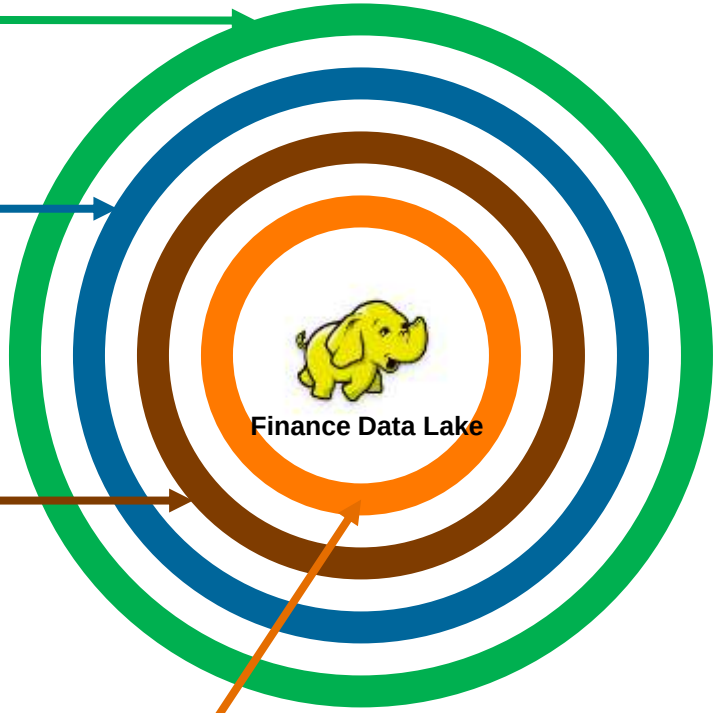
Security

Perimeter Level Security:
- Network Security (firewalls)
- Apache KNOX (Gateway – BI Tools)

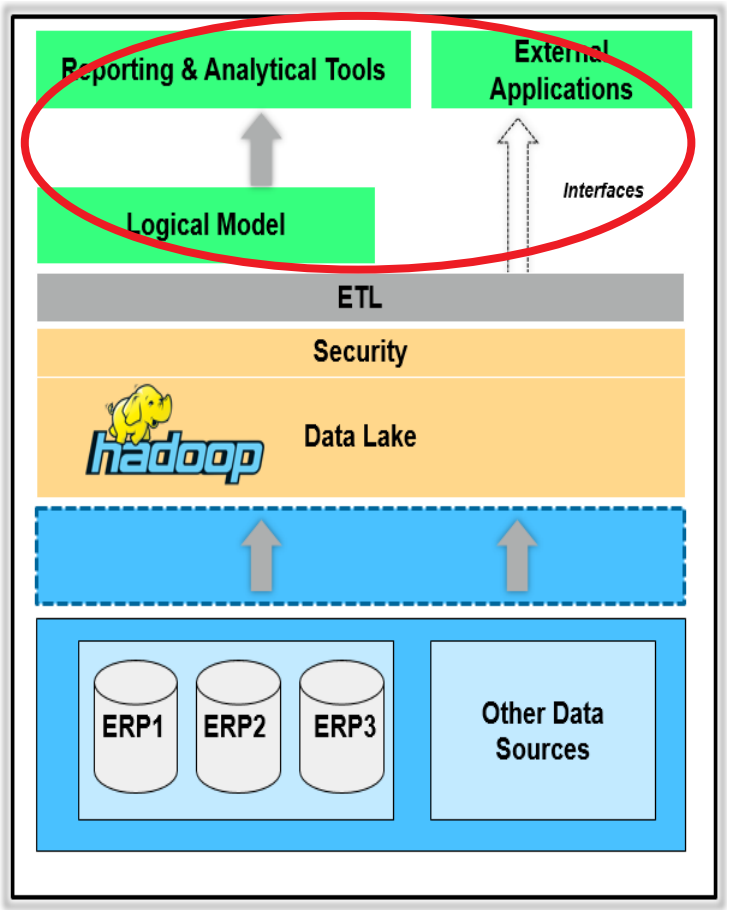
Access: (To Hadoop Cluster)
- Authentication
- Kerberos (Direct Access)

Data: (Protecting data in Cluster)
- Authorization
- Ranger
  - Role Based
  - Row/Column Level
- Encryption
- Data Masking

Audit/Administration
- Access Review
- Log Monitoring



Self-Service Discovery



Data Lake Platform

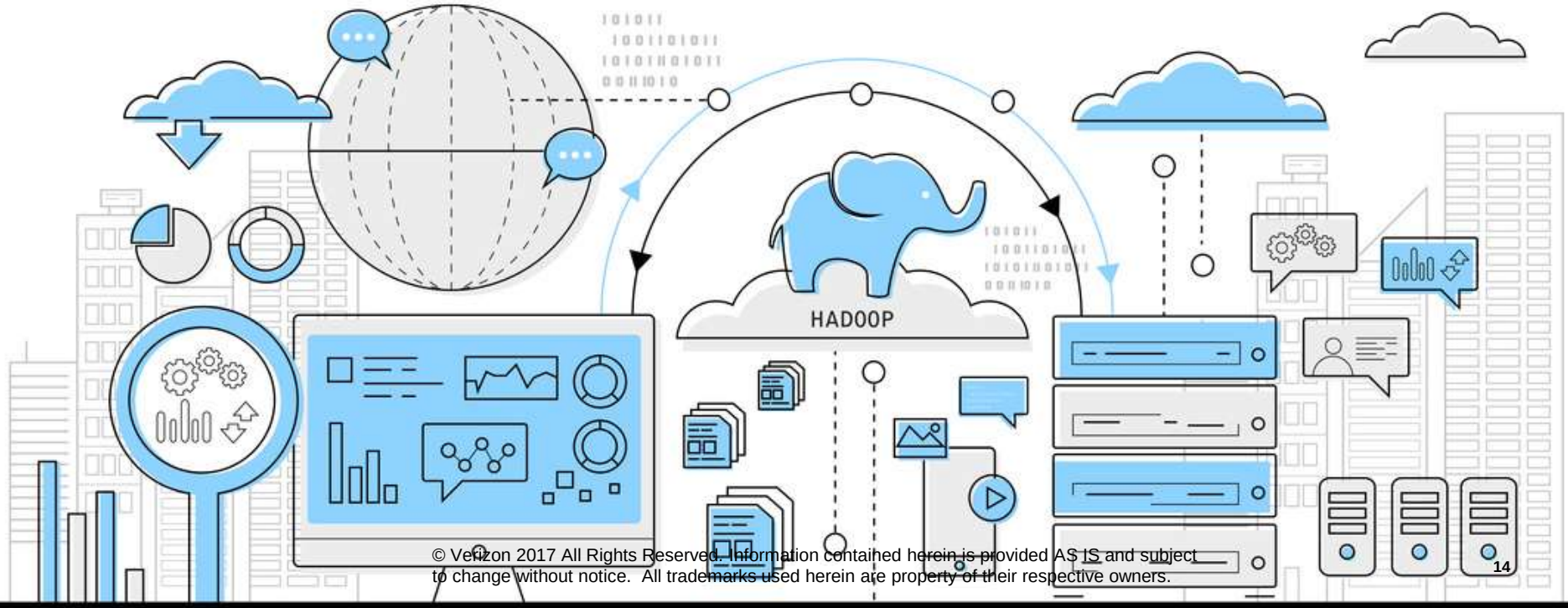
Explore	Model	Transform	Discover	Prepare & Share
relevant data	data to understand its potential	& enrich data to make it ready for analysis	powerful New insights	insights for enterprise leverage

Discovering the True Potential of Big Data

verizon

Takeaways

- **Data lake based on Hadoop big data platform is the right choice for self service discovery & analytics**
- **Adopt an Agile mindset in the implementation**
- **Evolve the architecture with the right tools for the right job**



Thank You

verizon