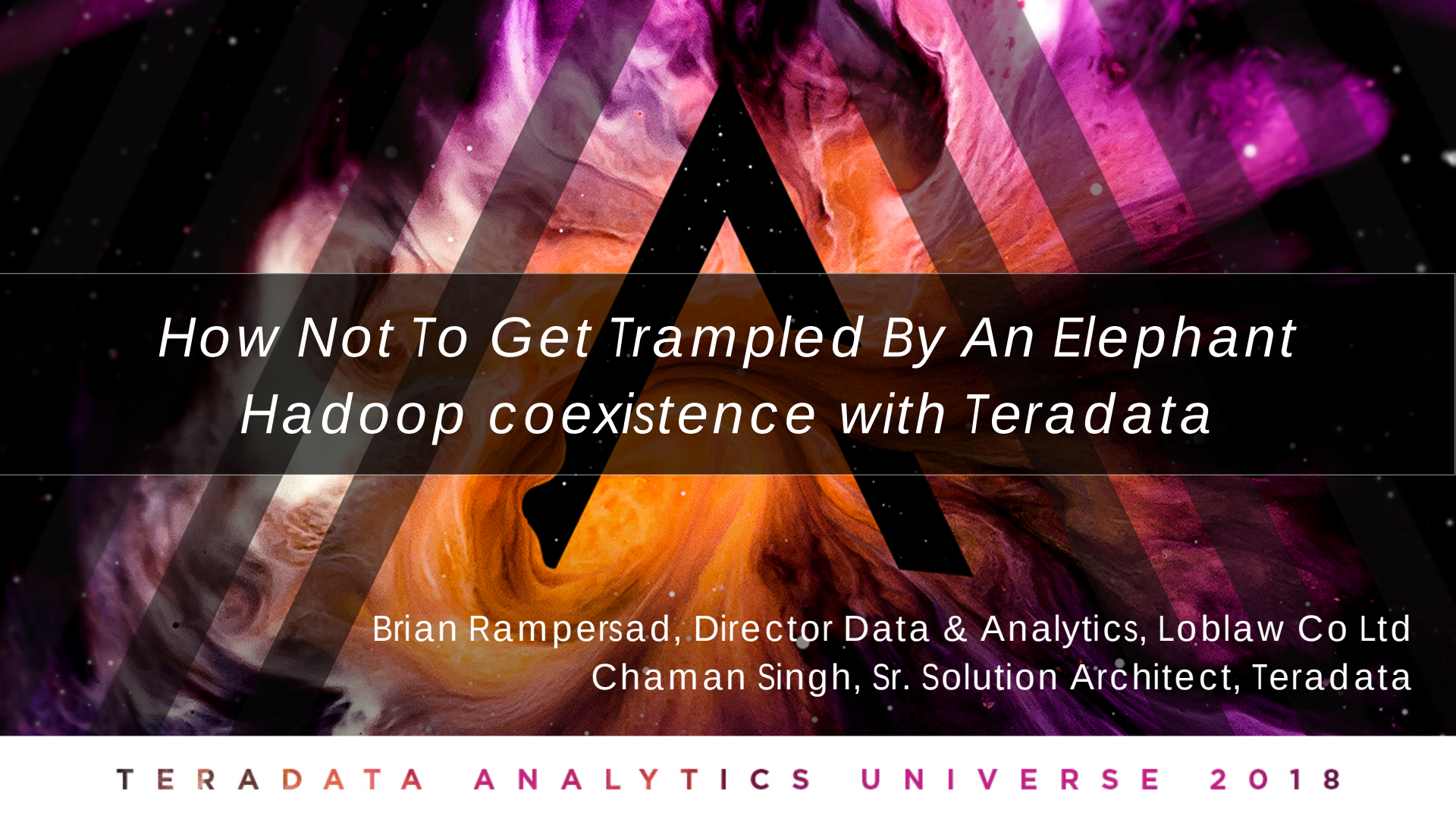


How Not To Get Trampled By An Elephant
Hadoop coexistence with Teradata

Brian Rampersad, Director Data & Analytics, Loblaw Co Ltd
Chaman Singh, Sr. Solution Architect, Teradata

Agenda

TERADATA
ANALYTICS UNIVERSE

- Who are we and why are we here?
- Introduction to Loblaw Companies Limited
- Program Objectives
- Program Approach
- Platform and Architecture
- Patterns and Design
- Optimization
- DR
- Summary

Overview of Loblaw Companies Limited *2017*

Nearly

200,000 employees

Canada's **food & pharmacy leader**

$46,385 billion
in revenue

$3,852 million
in EBITDA

$45,384 million
in retail segment sales

2,500
corporate, franchised and Associate-owned locations

The nation's

LARGEST

retailer

Best in food, health and beauty

Trusted Canadian brands

High-quality private label alternatives

Four of the top ten brands in Canada
(President's Choice, no name, Farmer's Market, Life)

Unique and superior products that offer better value

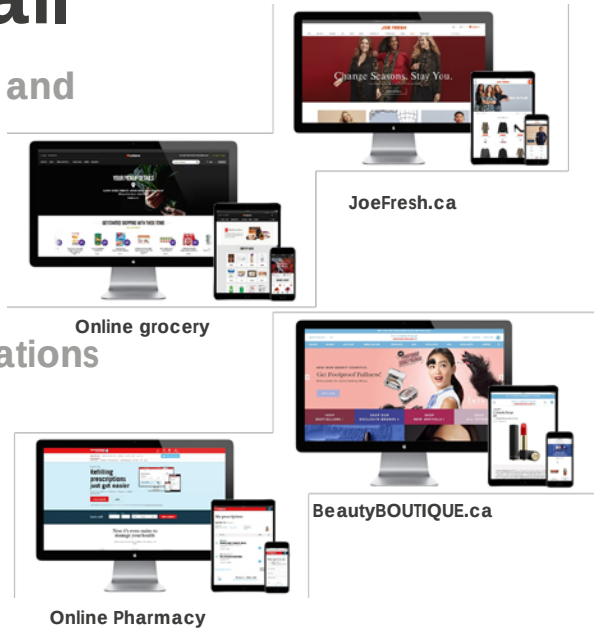
Everyday digital retail

Making shopping easier and more convenient

Online grocery, beauty, apparel and healthcare

200+ Click & Collect locations

Instacart partnership

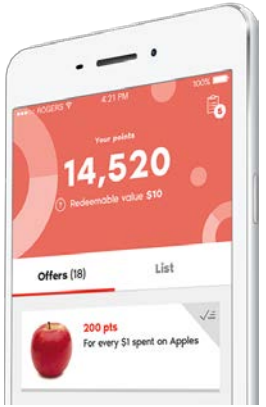


Payments and Rewards

Canada's favourite loyalty programs (Shoppers Optimum and PC Plus) have come together to create the PC Optimum program

Personalized offers for 19 million members

Earn more points with a PC Financial Mastercard



Our Program Objective

TERADATA ANALYTICS UNIVERSE

Enable an on premise Hadoop Data Integration Hub (DIH) to serve as a operational data repository.

This environment needed to support:

- Data Ingestion
- Data Publication
- Co-existence with Teradata
- Analytical workloads
- Disaster Recovery

How did we approach the work

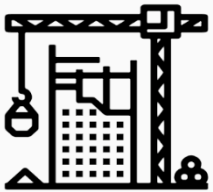
TERADATA ANALYTICS UNIVERSE

Build an Ecosystem

- Business requirements
- Data Lake architecture
- Choosing the right platform
- Data Integration Hub architecture
- Phased delivery approach
- Data lake implementation journey
- Performance optimization
- Operational monitoring

Platforms & Tools

- Hive
- Choosing the right tools
- ETL design- Patterns
- Performance Optimization
- Hadoop DR



Business Requirements

TERADATA
ANALYTICS UNIVERSE

- House structured or unstructured data
- Publish source Metadata
- High data availability
- Data ingestion supporting business SLAs
- High source data volumes
- Support for business sources analytical data processing
- Publish both Raw and Curated data
- Data Protection - PII Data Tokenization

Plan for slow ride

TERADATA ANALYTICS UNIVERSE

Crawl
Initial phase

Walk
Operationalize Hadoop

Run
Advanced Analytics

Prepare for both success and failure

TERADATA ANALYTICS UNIVERSE

- What data belongs in Hadoop
- Plans for success
- Set small achievable goals
- Learn from failures fast
- Work closely with the business
- Design it right– Data Lake architecture
- Develop with experience
- Implement using existing process
- Have additional hardware resources

Things to Keep in mind - before you start

- Hadoop/Hive is not Teradata
- POC to prove Hadoop and tools satisfy your use cases
- No one size fit all in Hadoop
- Power of Hadoop is within
- Choose technologies proven to work with Hadoop
- Use HDFS directly wherever you can
- Only reason you need traditional ETL tools is because you are used to it

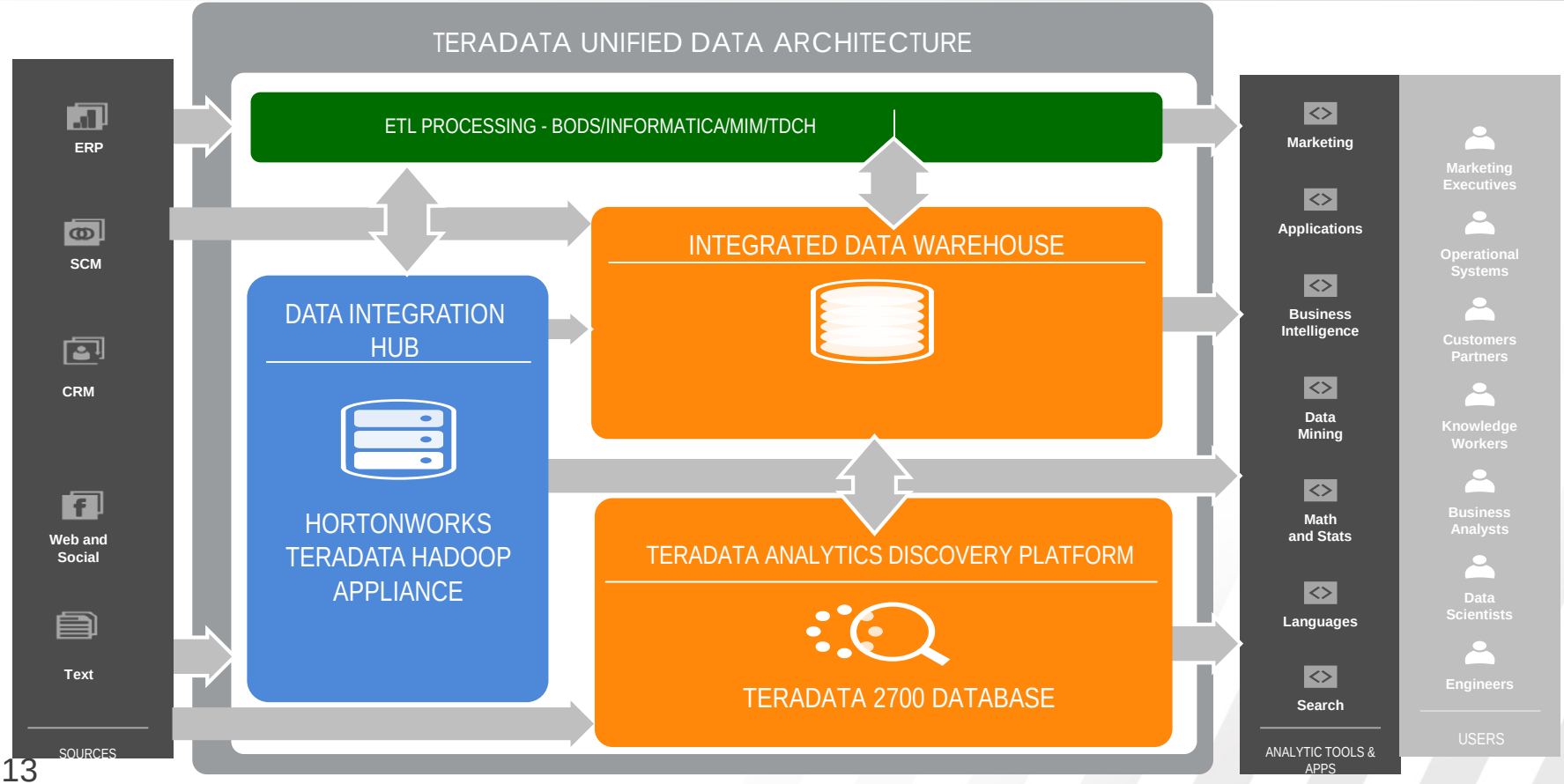
Choosing the right platform

TERADATA ANALYTICS UNIVERSE

	Analytics on Hive (Batch Processing)	Analytics on Teradata(Interactive)	Analytics on Hive(Interactive)	Analytics on Teradata(Interactive
Data Volume	High	Low	High	Low to High
User Experience SLA	Low	High	High	High
Data Freshness	Day-1	Day-1	Day-0	Day-1
Concurrency	N/A	> 40	< 40	> 40
Data Migration (Hadoop->TD)	N/A	Low	N/A	Low to High
Data Processing complexity	Low	Low	Low	High

TERADATA ANALYTICAL ECOSYSTEM

TERADATA ANALYTICS UNIVERSE



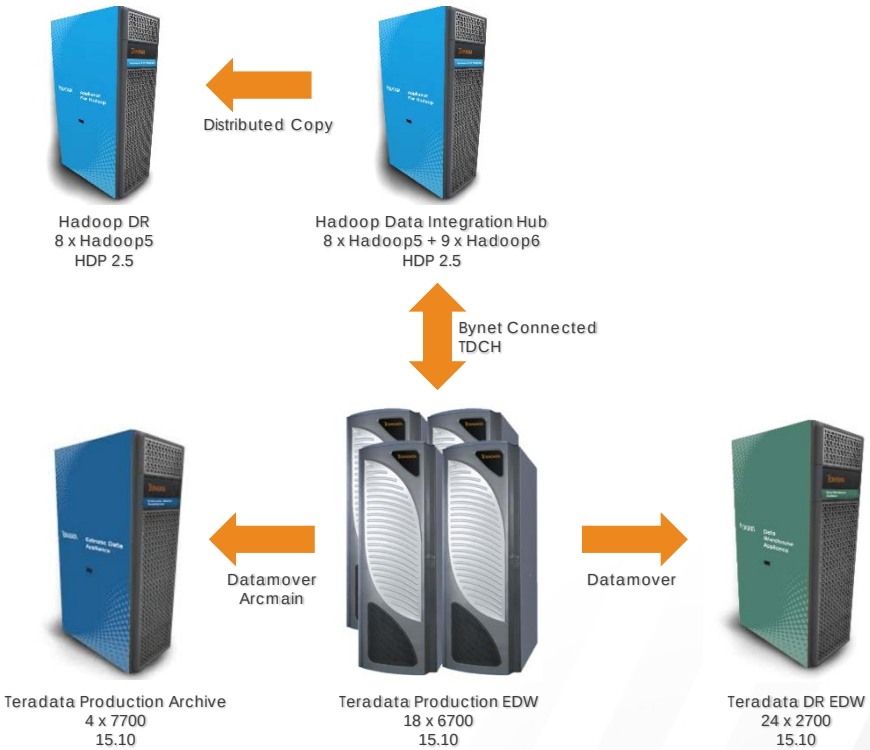
Data Lake Architecture

TERADATA ANALYTICS UNIVERSE

- Invest in reusable frameworks
- Keep architecture simple
- Minimize the data movement
- Bring analytics closer to the source
- Bring everything to data lake
- Push ETL processing down to Hadoop
- Develop ETL/access design patterns
- POC the tools you choose
- HDFS security/Kerberize cluster
- Workload management
- Tokenize PPI/PHI data

Analytical Ecosystem @ Loblaw

TERADATA ANALYTICS UNIVERSE



python

R

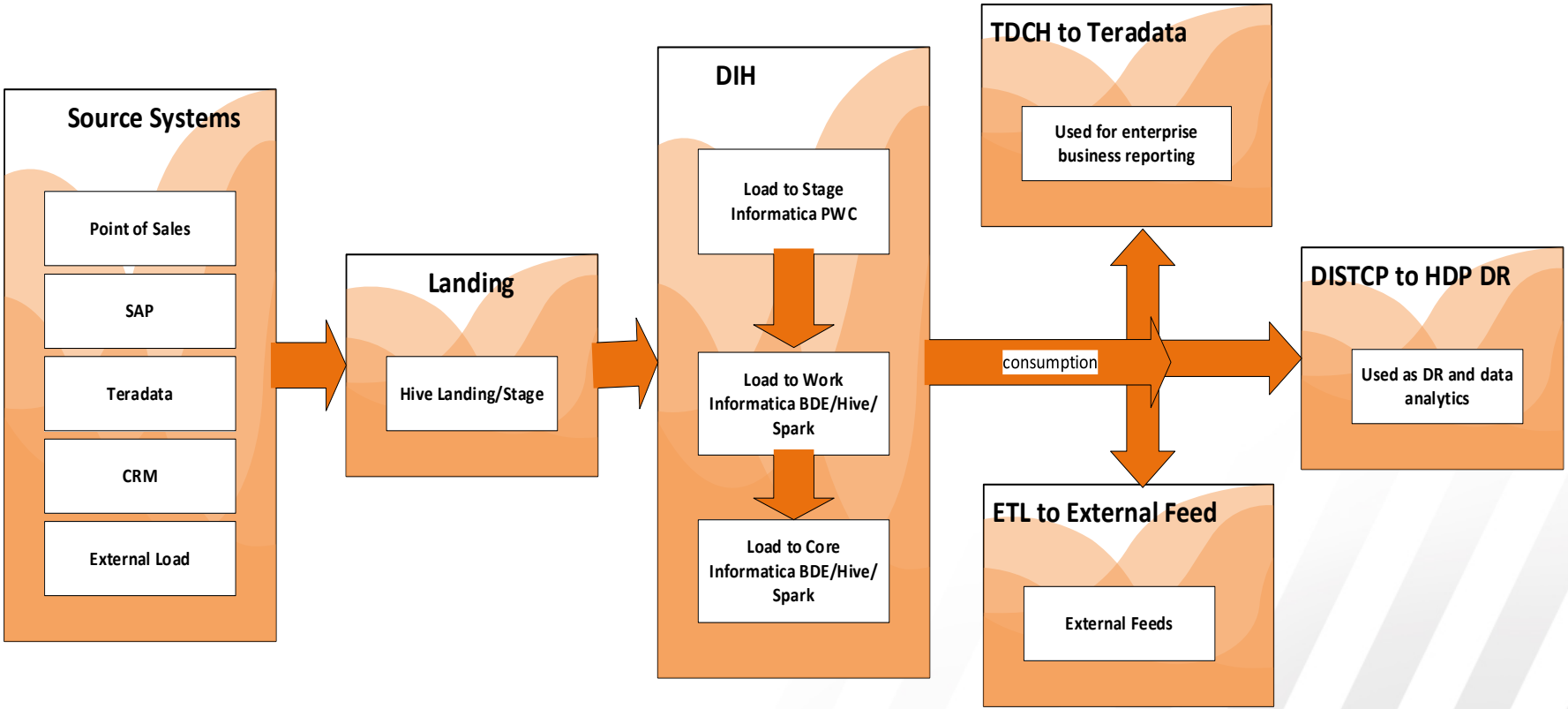
sas
THE POWER TO KNOW.

MicroStrategy

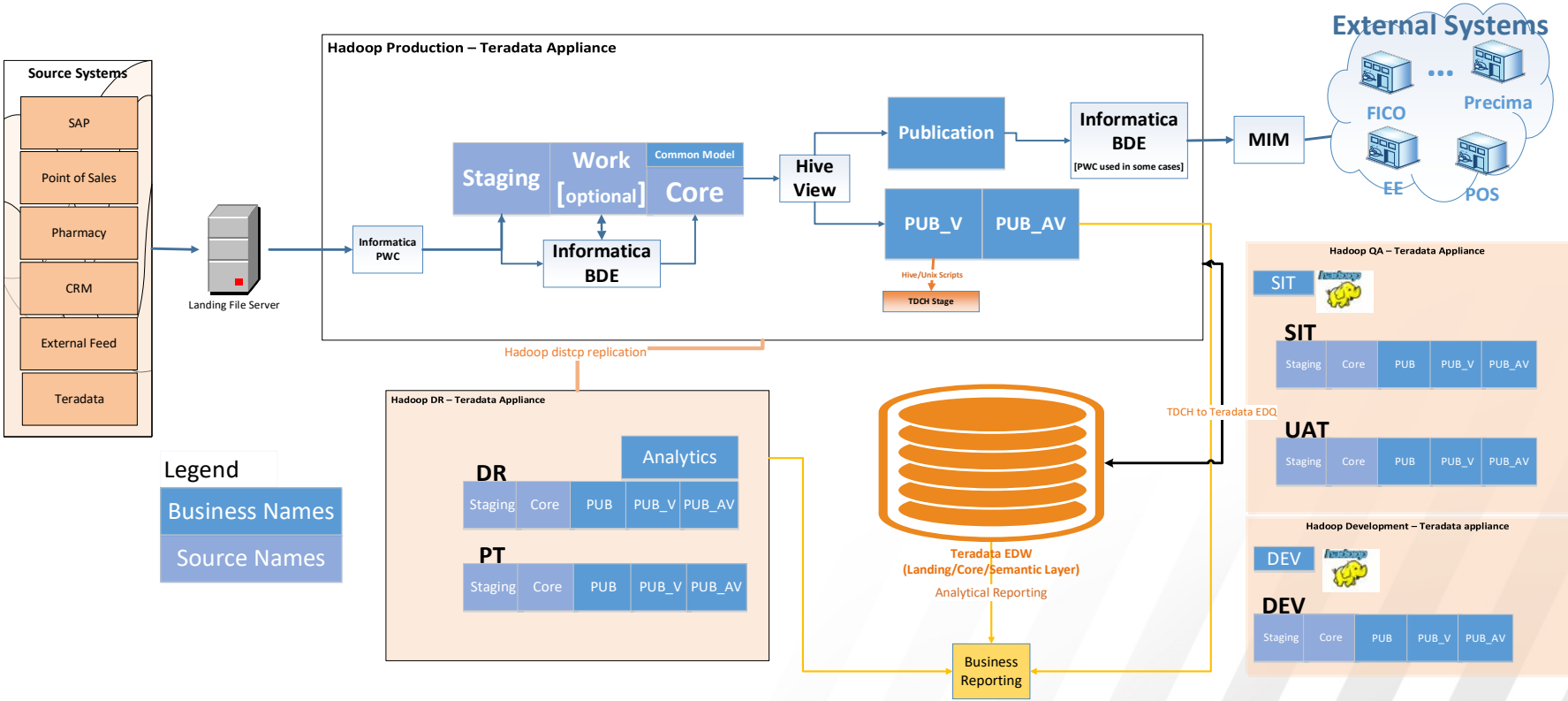
Power BI

Data Integration Hub architecture

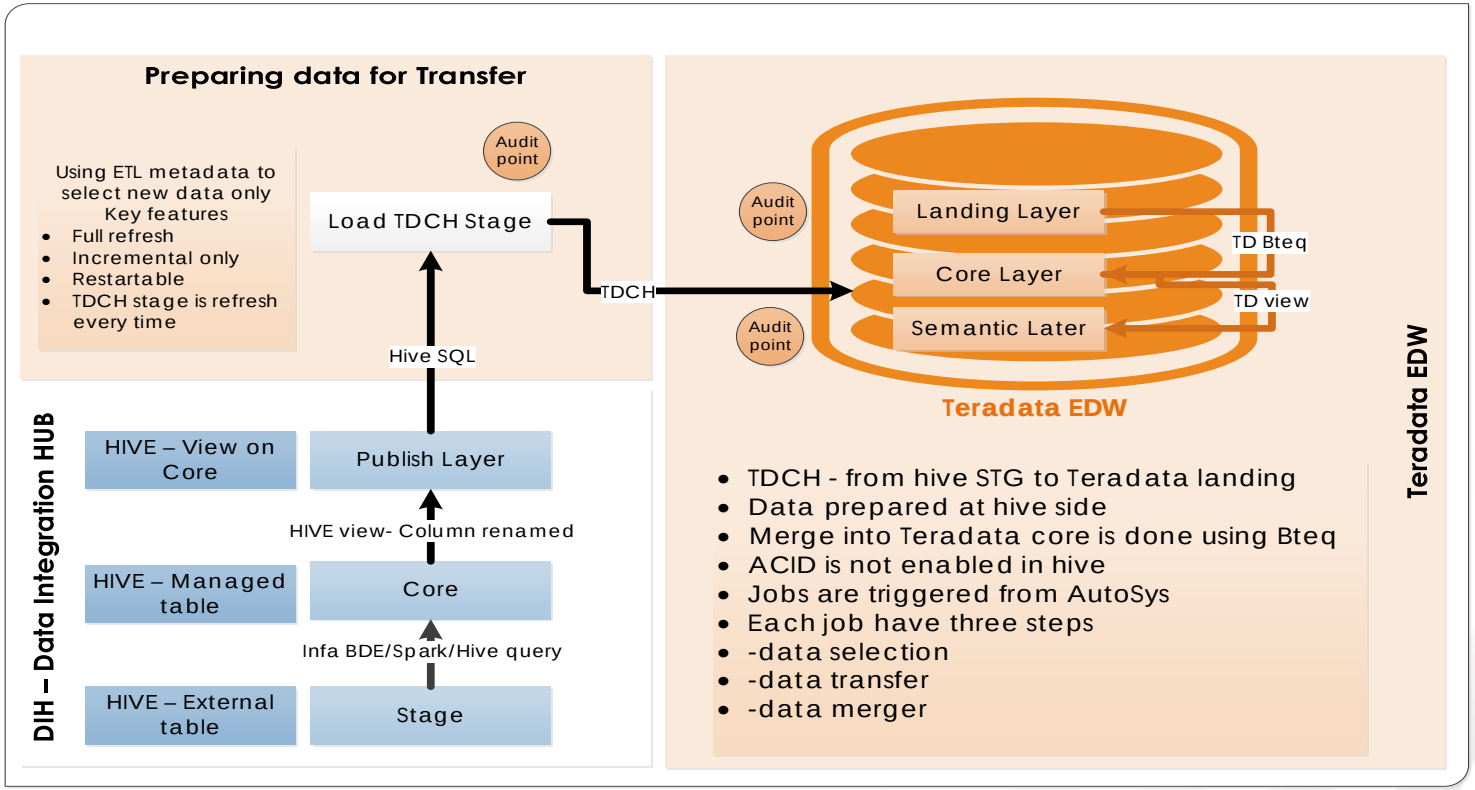
TERADATA ANALYTICS UNIVERSE



Data Integration Hub architecture



TDCH – Loading to Teradata



Choosing right tool for a job

TERADATA ANALYTICS UNIVERSE

Implementation Alternatives	Interface	Storage Platform	Processing Engine	Flexibility	Performance	Implementation Complexity	Streaming
Hive (MR/TEZ)	HiveQL	HDFS File/Hive/ Hbase*	Map Reduce, Tez	Low	Med	Low	No
Spark SQL	Scala	Any	Spark	High	High	Medium to High	Yes
Pig	Pig Latin	HDFS File/Hive	Map Reduce, Tez, Spark	Low	Med	Low to Medium	No
Hbase/Phoenix	Phoenix	Hbase	HBASE API	High	High	High	Yes
Custom Code	Scala, Java	Any	Any	High	High	High	Yes

Hive data modeling

TERADATA ANALYTICS UNIVERSE

- Hive object Naming standard
- HDFS directory naming standard
- Data modeling best practices
- DDL generator/deployment packages
- Partition considerations
- Hive is schema on read – design for that
- Common Publish layer for business(views)
- Materialize resource intensive processing
- Handle small files
- Insert only data grows fast- consider that
- Follow EDW standards
- Metadata is not optional
- Avoid excessive abbreviations

ETL Design patterns

TERADATA ANALYTICS UNIVERSE

- Calculate incoming data volume (initial and ongoing)
- Determine type of data(Master data or Transaction)
- Understand how data will be consumed
- Determine the required refresh frequency
  - Master Data – Full refresh
  - Master data – Snapshot(alternate for CDC)
  - Transaction data – Incremental everyday
- Choose right tool for job/design
- Data availability SLA – Source and Target

ETL design Pattern – Master data

TERADATA ANALYTICS UNIVERSE

- **Injection ETL - key points**
  - Full refresh
  - Business Effective Date/Load Date
  - DIH Core loaded with Record Create
  - Partition Hive tables for optimal retrieval

- **How to consume**
  - Apply previous date's filter to select latest row
  - Select row with max Business Effective Date/Timestamp for each business key. Only if options 1 does not work.
  - As of any date reporting is possible and easier to achieve

ETL design Pattern – Master data

TERADATA ANALYTICS UNIVERSE

- **Injection ETL - key points**
  - Incremental load only
  - Insert only
  - Rows has Business Effective Date
  - No existence check business key

- **How to consume**
  - Select row with max effective date
  - As of date reporting is not easily achievable but possible.

ETL design Pattern – Transaction data

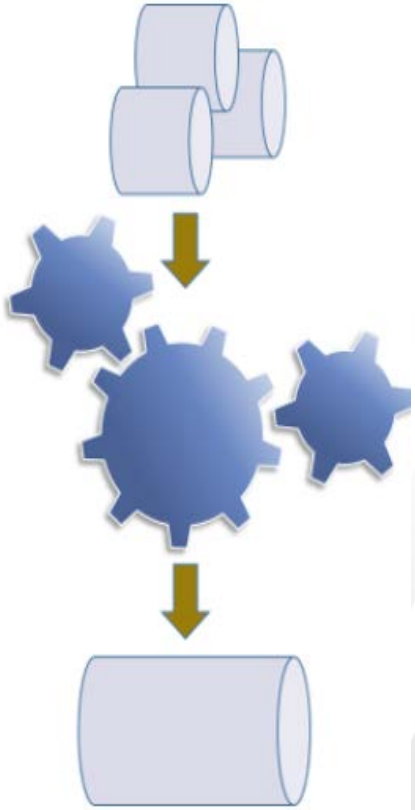
TERADATA ANALYTICS UNIVERSE

Injection ETL - key points

- Only incremental data is loaded to DIH Core
- Source provide the previous rows associated to that business key.
- ETL is insert only with no existence check
- Data is populated into DIH Core with Business Effective Date/Timestamp of when ETL

Data Consumption- key points

- Select row with max Business Effective Date/Timestamp for each business key.
- Create Hive materialize table to maintain the latest version of all business key
- As of date reporting is not easily achievable.



Performance Optimization

TERADATA
ANALYTICS UNIVERSE

- Consolidation in Hive or HDFS
- Push ETL processing to Hive
- Consume common consolidated data
- Consolidated data at lowest level
- Create hive aggregate tables for performance
- Use Spark for materialization
- Move aggregated data Teradata
- Key Business data integrated with Teradata

Hadoop DR Replication -DistCP

TERADATA ANALYTICS UNIVERSE

Pros

- Custom build DisctCp solution
- Unix scripting
- Automated using Autosys
- Perform well
- Easily portable

Cons

- Fail frequency during data load
- Difficult to investigate
- No inbuild metadata sync-up functionality
- Missing file failure

Metadata Sync-up :

Customized solutions using scripting.

It works but **difficult to maintain** and **frequent failures**

Operational Monitoring

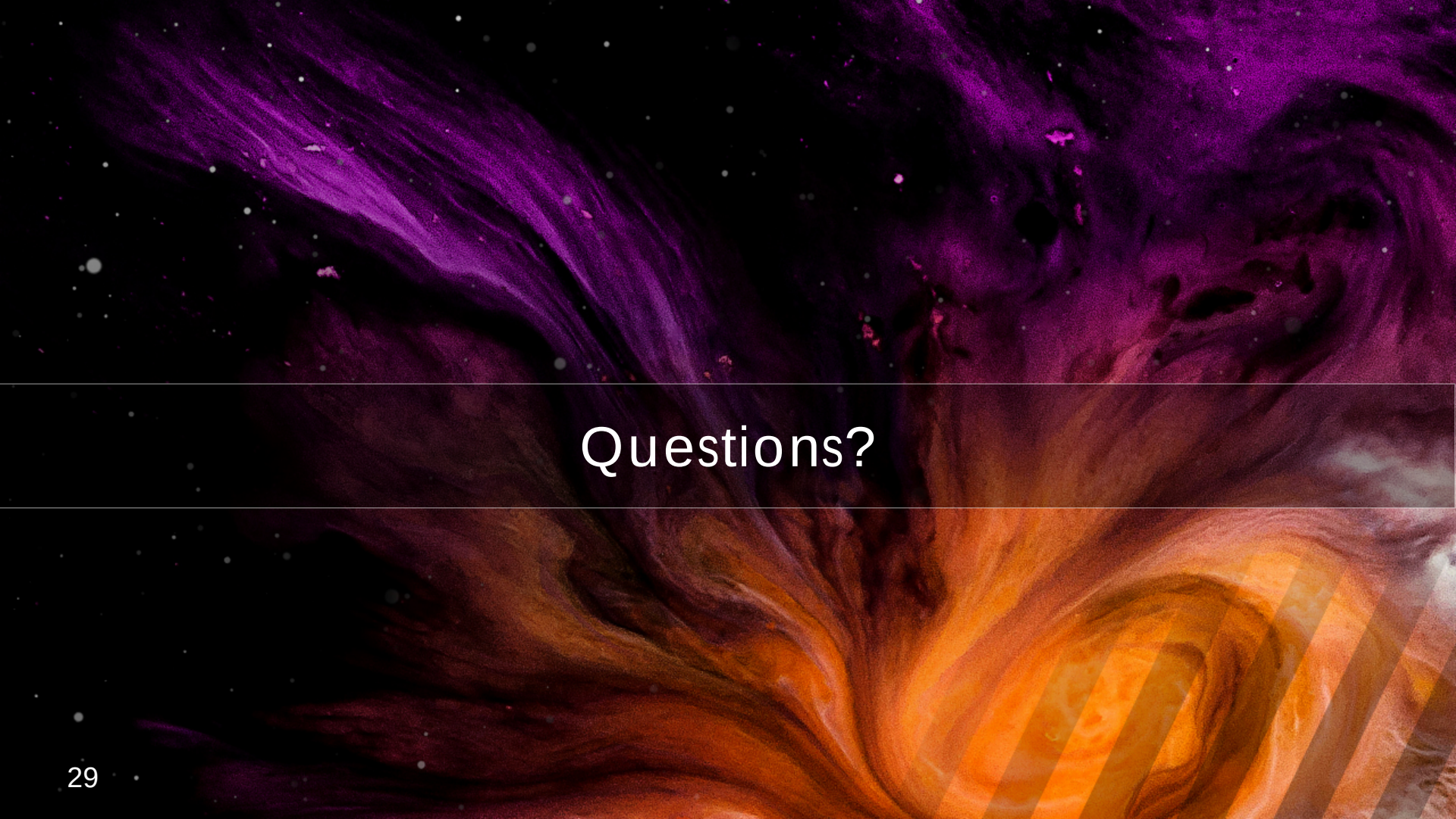
TERADATA
ANALYTICS UNIVERSE

- Support intensive –Hadoop admin team
- Server level error are frequent – Business communication
- Manual Hadoop cluster metrics
- Manual ETL run metadata metrics
- Business expects it work as Teradata – which is hard
- Workload is minimum limit not maximum
- Regular maintenance activities
- Regular profiling of data/analyze table
- Lots of delete and reload
- Installing new tools

Few more things to consider

TERADATA ANALYTICS UNIVERSE

- TDCH or Sqoop
- Hive DDL Generator and convertor
- Lower environments
- Tool installation and POC
- Data Modeling tools
- Data types in Hive
- Maintaining history in Hive/HDFS
- CDC in hive
- Workload management


Questions?

Thank You!

Rate This Session # 0900

with the Teradata Analytics Universe Mobile App

Questions/Comments

Email: brian.rampersad@loblaw.ca & chaman.singh@teradata.com