Hadoop Overview for Managers

PART 2

Outline

PART 1

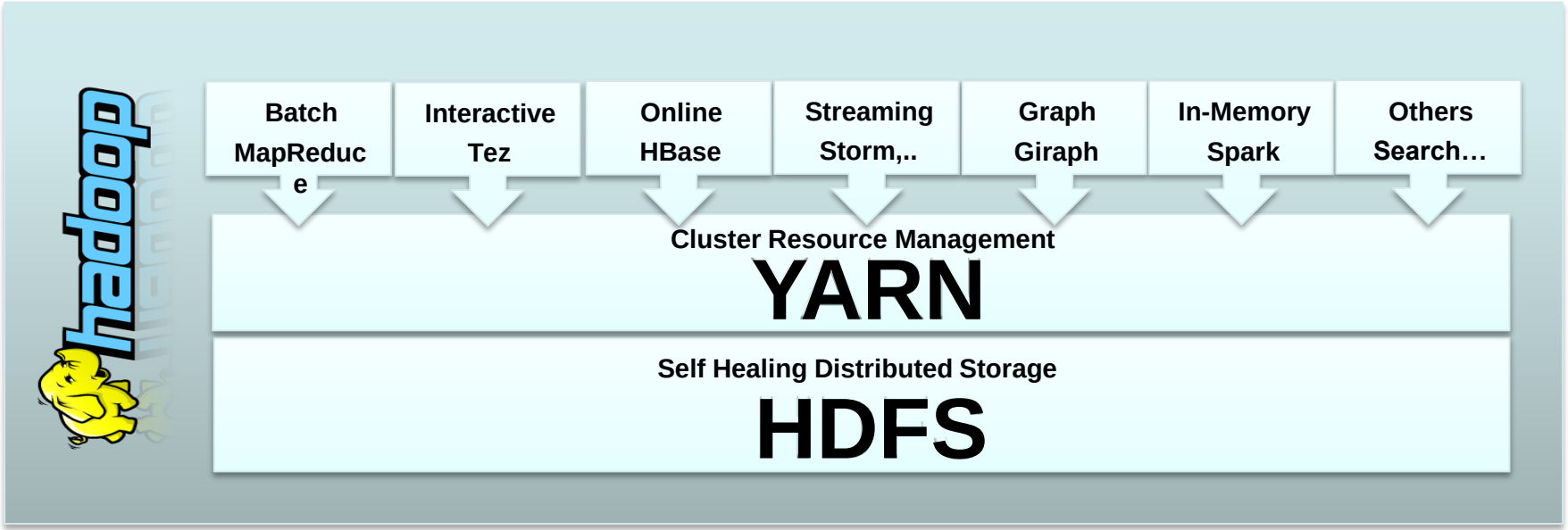
- Hadoop Overview
  - Traditional Computing Systems and Limitations
  - Why Big Data?
  - Why Hadoop?
  - Hadoop Basic Concepts
  - Where Hadoop fits in the Enterprise
- Hadoop Architecture
  - Building blocks
  - HDFS
  - Demo
- **Break**

PART 2

- YARN Architecture
  - Yarn Overview
  - MapReduce
  - Demo
- Tools and technology for Hadoop ecosystem
- Hadoop Real Life Use Cases
- Establishing a Big Data Center of Excellence
  - Justifying business value for your organization
  - Challenges on building a production solution
  - Recommended organizational structure
  - Best Practices: Steps to effectively deploy Hadoop
- Recap and Q&A

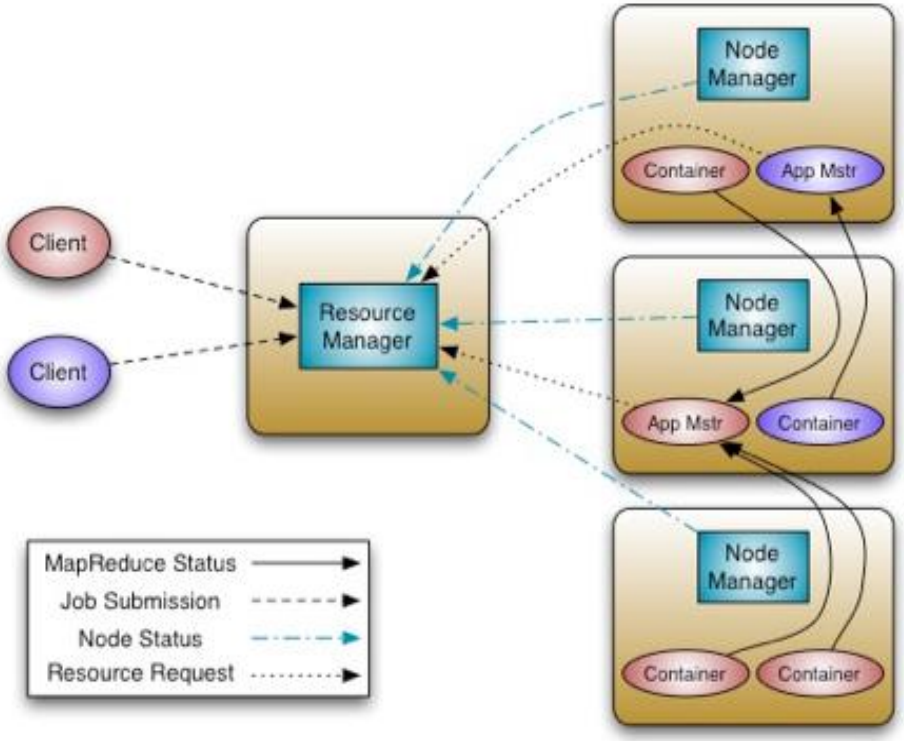
YARN OVERVIEW

Hadoop YARN



- Hadoop 2.X is now architected to handle various types of workloads Batch, interactive, Streaming and others

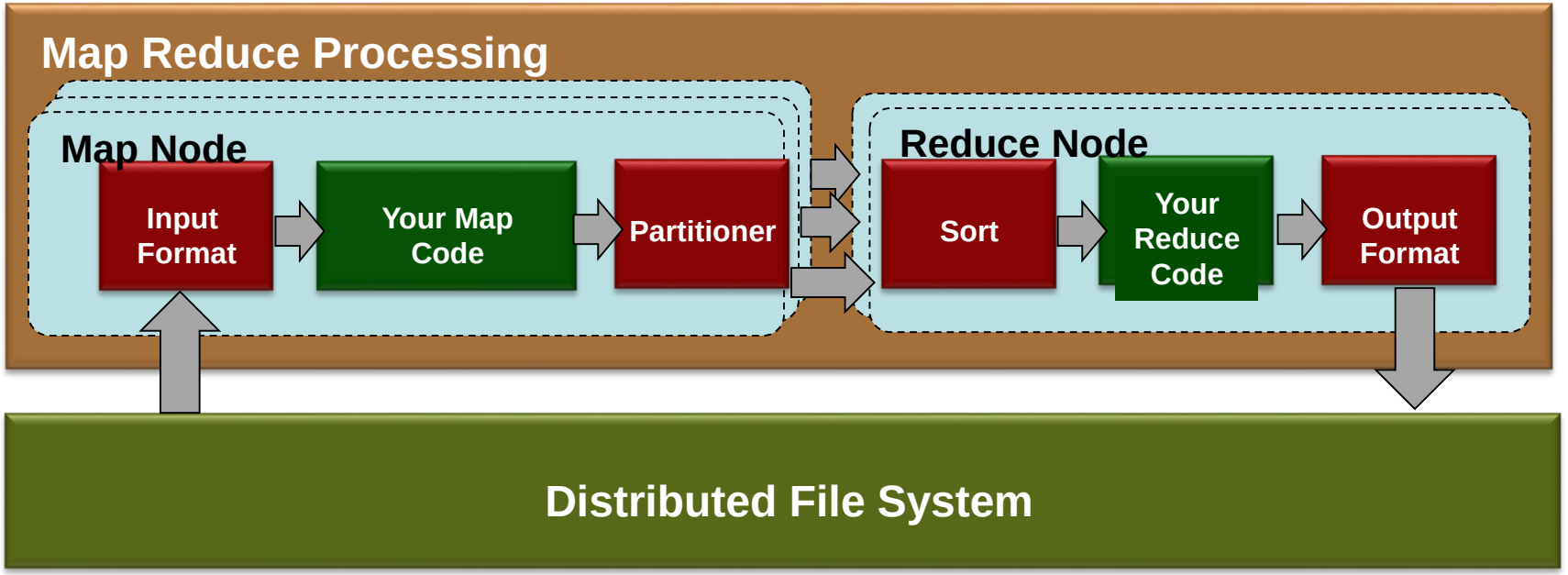
YARN Overview



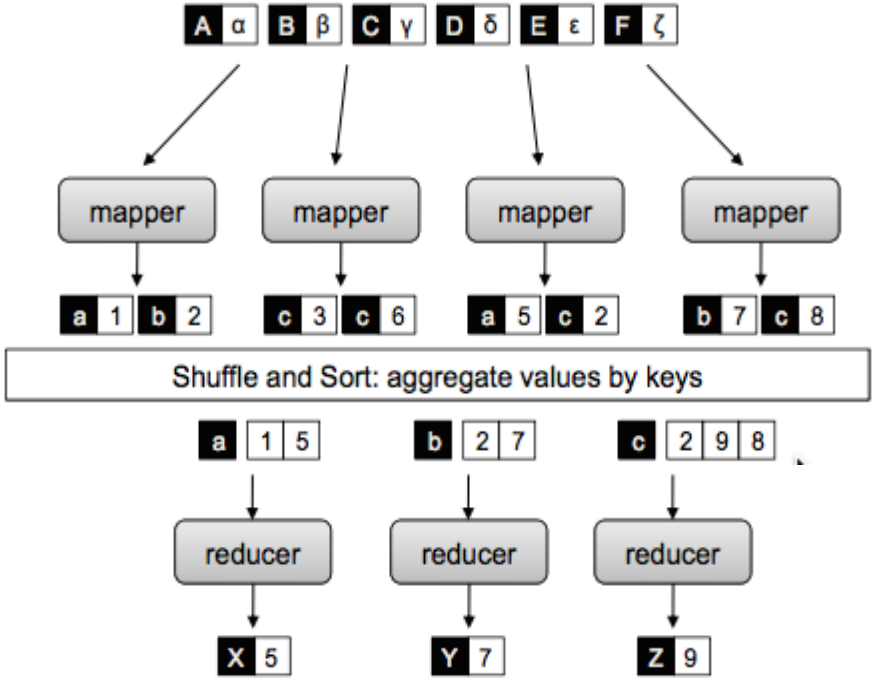
- Resource Manager
  - Manages Schedules
- Node Manager
  - Containers for map/reduce or any other application
- Application Master
  - Distributed and can run on any node

Courtesy: http://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html

Map Reduce Architecture

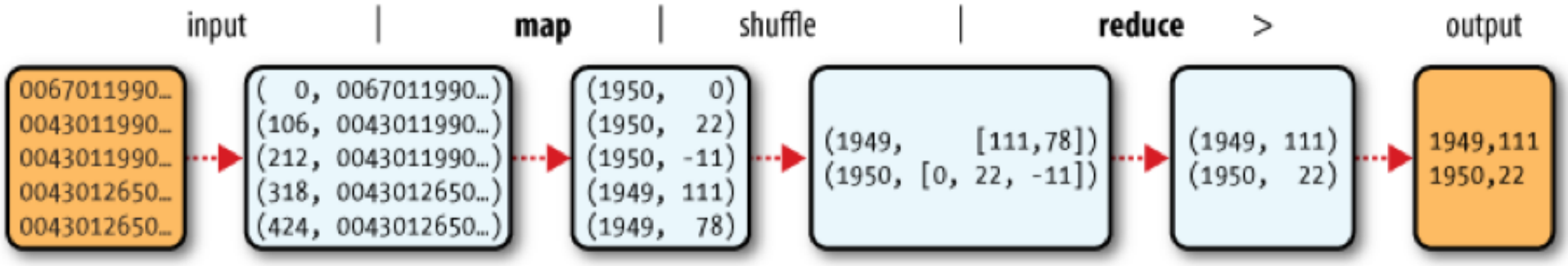


Map Reduce Overview



Source: Data-Intensive Text Processing with MapReduce Jimmy Lin and Chris Dyer

Example Logical flow




```
0067011990999991950051507004...9999999N9+00001+99999999999...
0043011990999991950051512004...9999999N9+00221+99999999999...
0043011990999991950051518004...9999999N9-00111+99999999999...
0043012650999991949032412004...0500001N9+01111+99999999999...
0043012650999991949032418004...0500001N9+00781+99999999999...
```

Source: Hadoop, The Definitive Guide

Another example: WordCount

- Large file contains unknown number of words
- Generate a list of unique words along with a count of how many times they occur

- Example:

Betty Botter bought a bit of butter

The butter Betty Botter bought was a bit bitter

And made her batter bitter

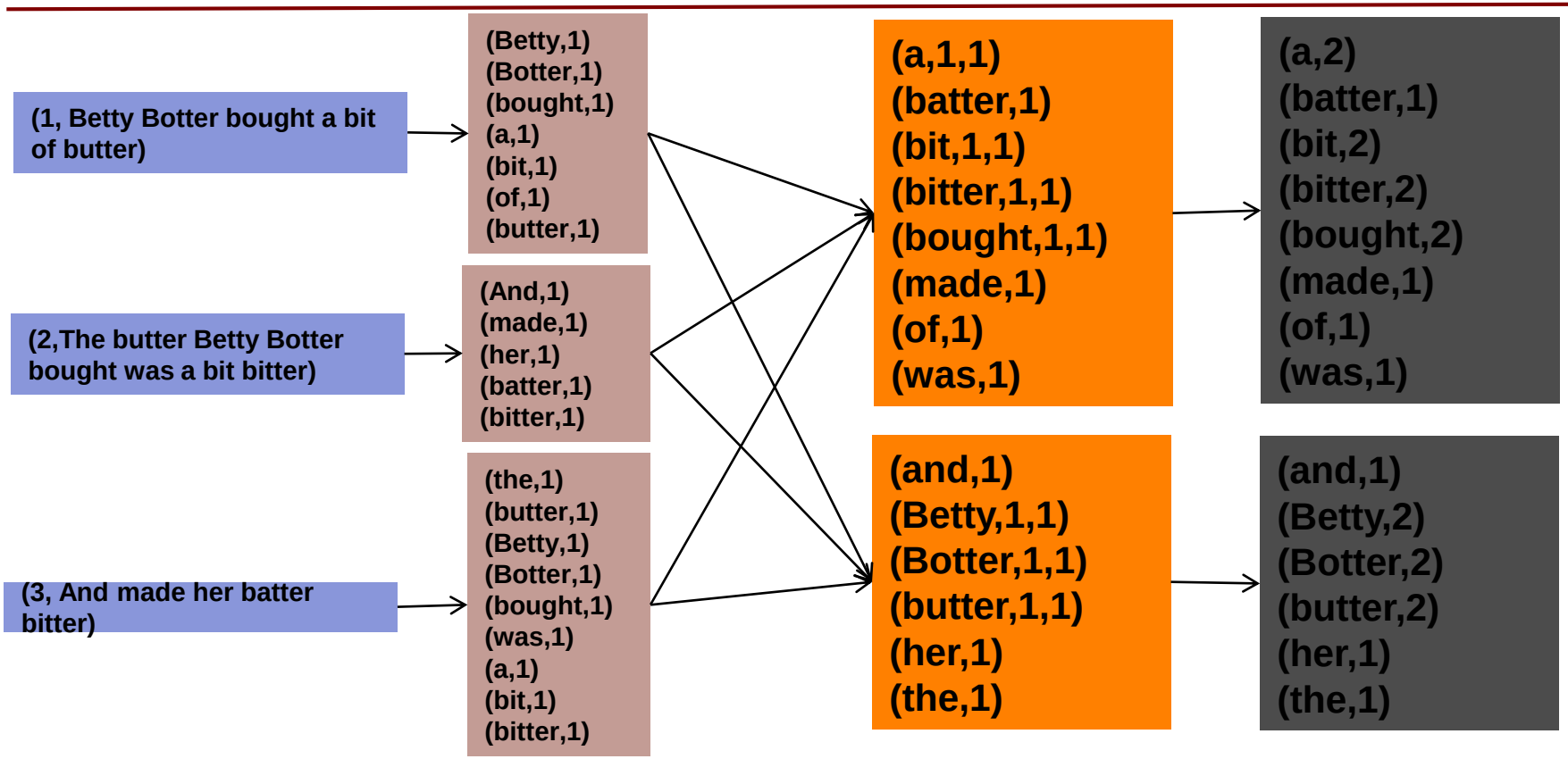
Word Count Example (Pseudo code)

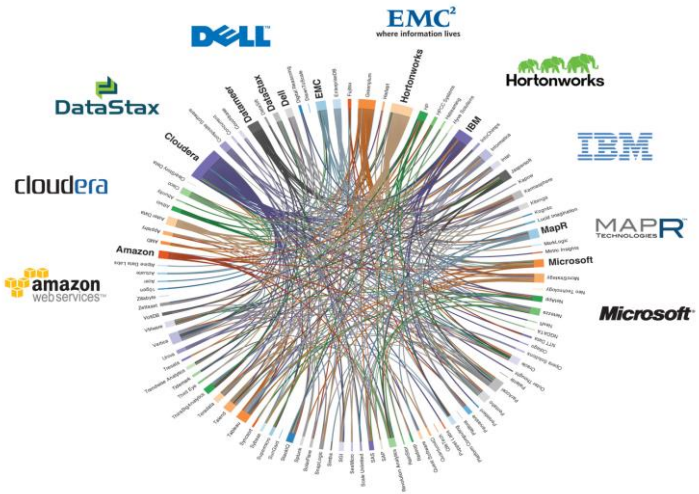
ZALONI

```
map(String input_key, String input_value):
  // input_key: document name
  // input_value: document contents
  for each word w in input_value:
    emit(w, 1);

reduce(String output_key, Iterator<int>
intermediate_values):
  // output_key: a word
  // output_values: a list of counts
  int result = 0;
  for each v in intermediate_values:
    result += v;
  emit(output_key, result);
```

WordCount

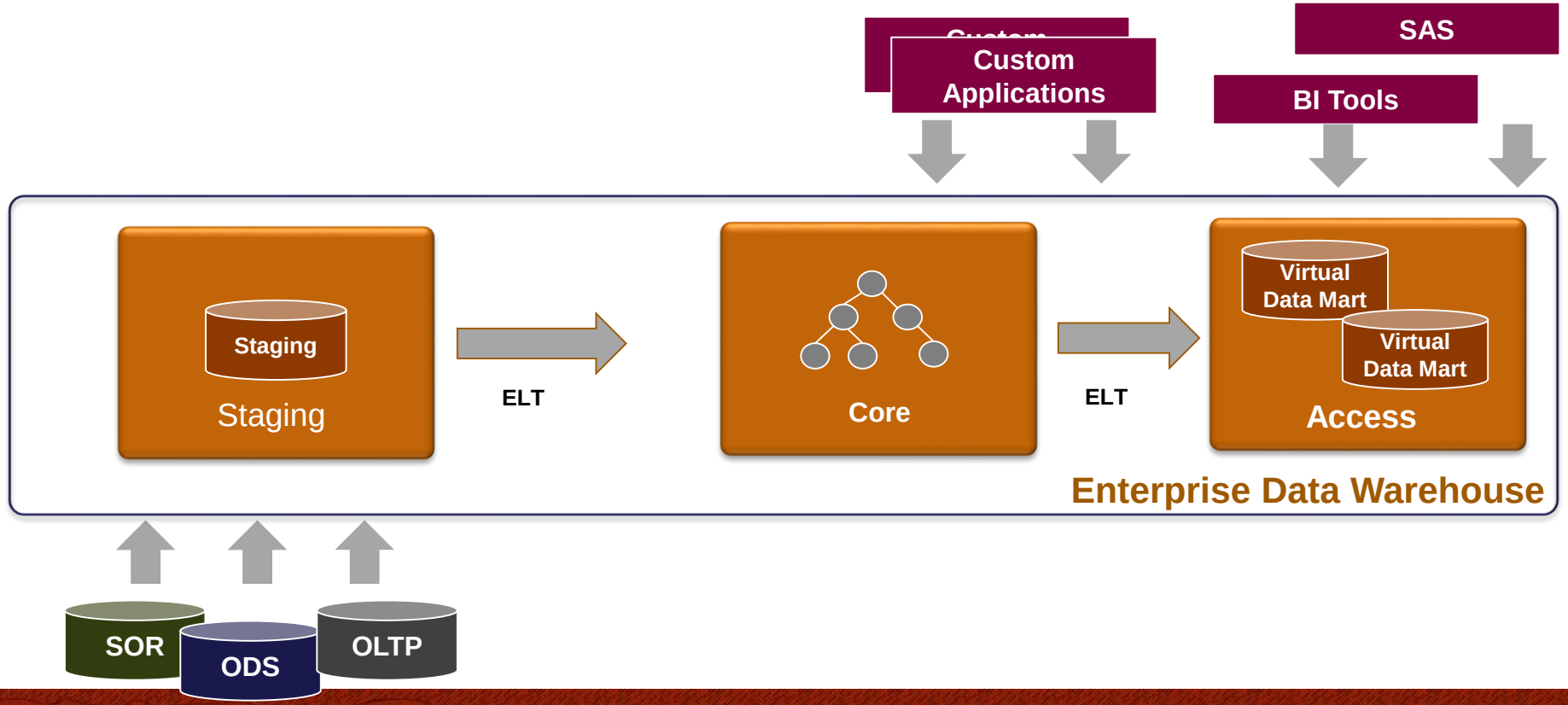




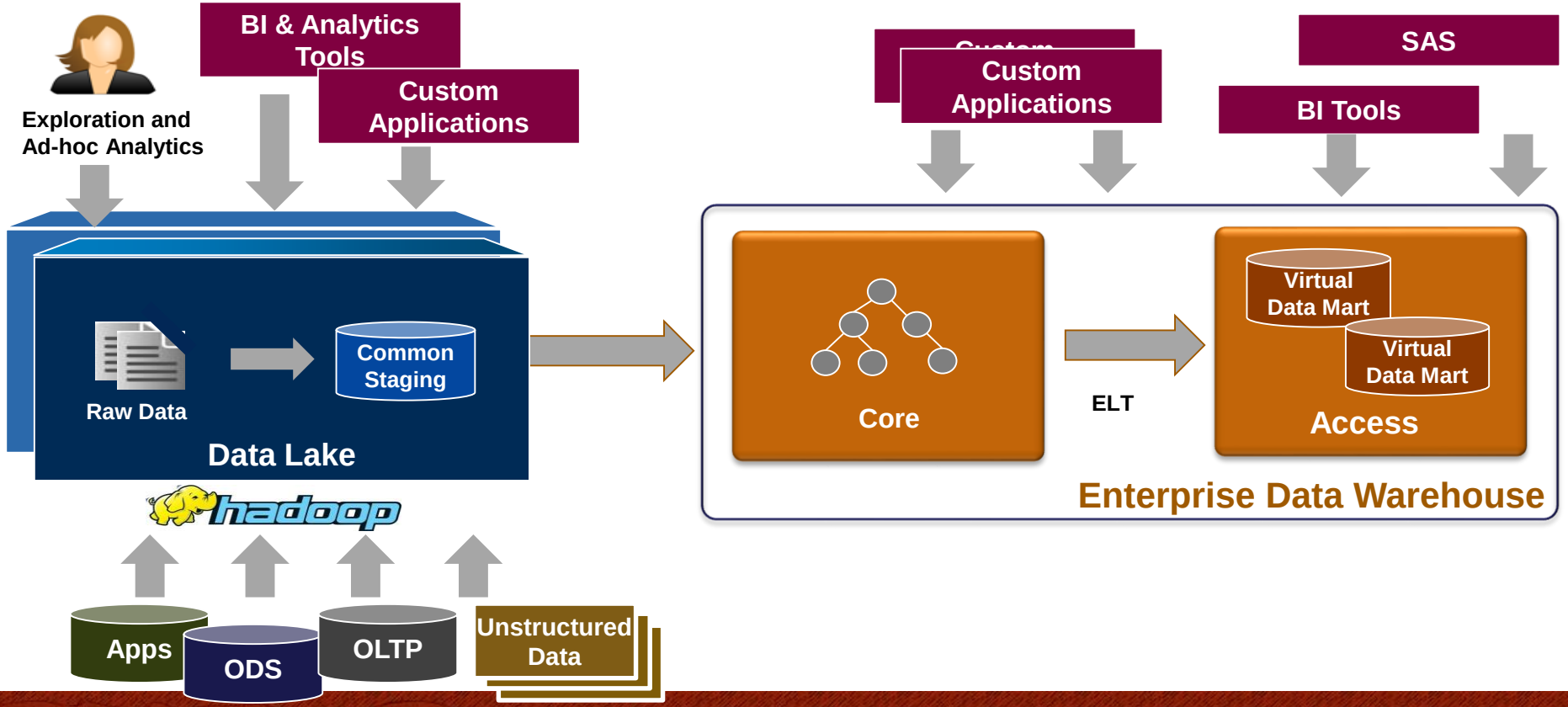
HADOOP IN THE ENTERPRISE

Image courtesy Datameer

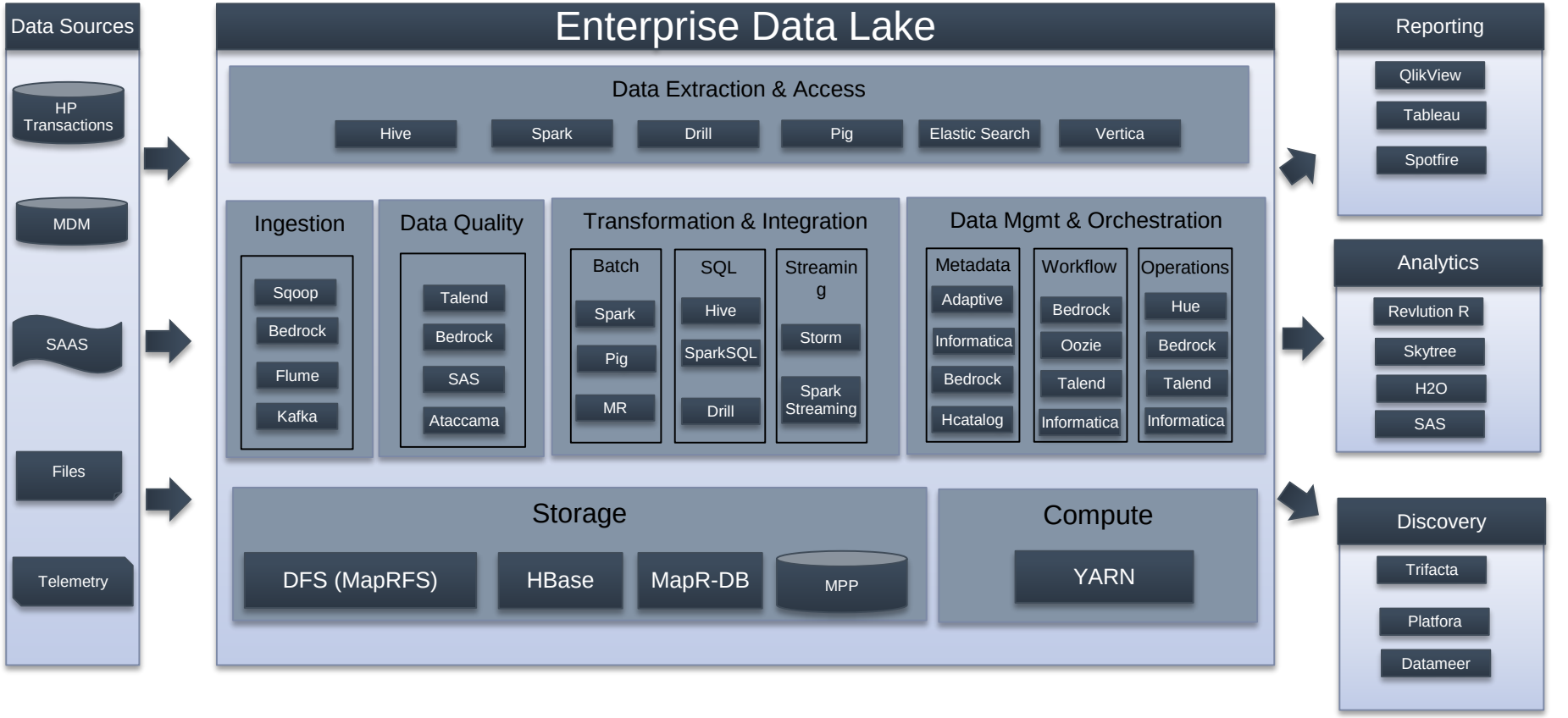
Traditional Data Platforms



Enterprise Data Lake



Hadoop: Foundation for Innovation



Nature of Hadoop-able Problems

- Complex data
- Multiple data sources
- Lots of it

Nature of Analysis

- Batch Processing
- Parallel Execution
- Data in distributed file system and computation close to data

Analysis Applications

- Text mining
- Risk Assessment
- Graph Analysis
- Pattern Recognition
- Sentiment Analysis
- Collaborative Filtering
- Prediction Models

Hadoop Use Cases

- Yahoo!: Social Graph Analysis
- VISA: Large Scale Transaction Analysis
- China Mobile: Data Mining Platform for Telecom Industry
- JP Morgan Chase: Data Processing for Financial Services
- eHarmony: Matchmaking in the Hadoop Cloud
- Rackspace: Cross Data Center Log Processing
- Visible Technologies: Real-Time Business Intelligence
- Booz Allen Hamilton: Biometric Analysis for DHS
- General Electric: Sentiment Analysis powered by Hadoop
- Bank of America: Call center analytics, Quantitative Risk

- http://wiki.apache.org/hadoop/PoweredBy

Hadoop Adoption in the Enterprise

ZALONI

HADOOP USE CASES

Case Study

HADOOP ETL OFFLOAD

Business Challenges

- Current EDW platform is used more for ETL processing than data warehousing and business intelligence
- Capacity Exhaustion as new Datasets need to be handled
- High turn around time of billed revenue extract to downstream data marts owing to limiting capacity on EDW platform
- EDW platform enhancement comes with significant high OpEx cost
- High risk backup, archival and restore processes which is manual

Business Objectives

ZALONI

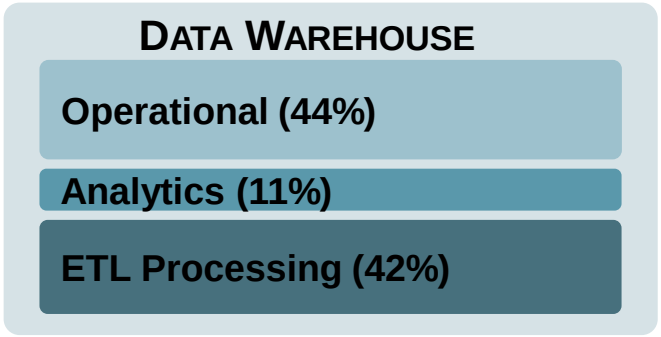
- Reduce Total Cost of Ownership (TCO) of the EDW
- Preserve current Business Logic while making the transition to Hadoop a sustainable and manageable platform.
- Shorter Time-to-Insight - Reduce turn around time of billed revenue extract to downstream data marts owing to limiting capacity on EDW platform
- Longer data retention of historical data for enabling analytics and data mining - Enterprise Data Archive
- Ad-hoc Analytics, Faster On-boarding of New Datasets
- Automate backup, archival and restore processes

Business case for ETL Offload

ZALONI

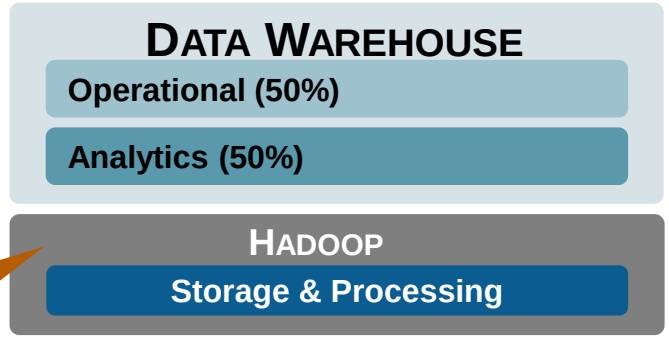
The Situation

- Many EDWs are at capacity
- Running out of budget before running out of relevant data
- Older data archived “in the dark”, not available for exploration



The Solution

- **Hadoop for data storage and processing: parse, cleanse, apply structure and transform**
- **Free EDW for valuable queries**
- **Retain all data for analysis!**

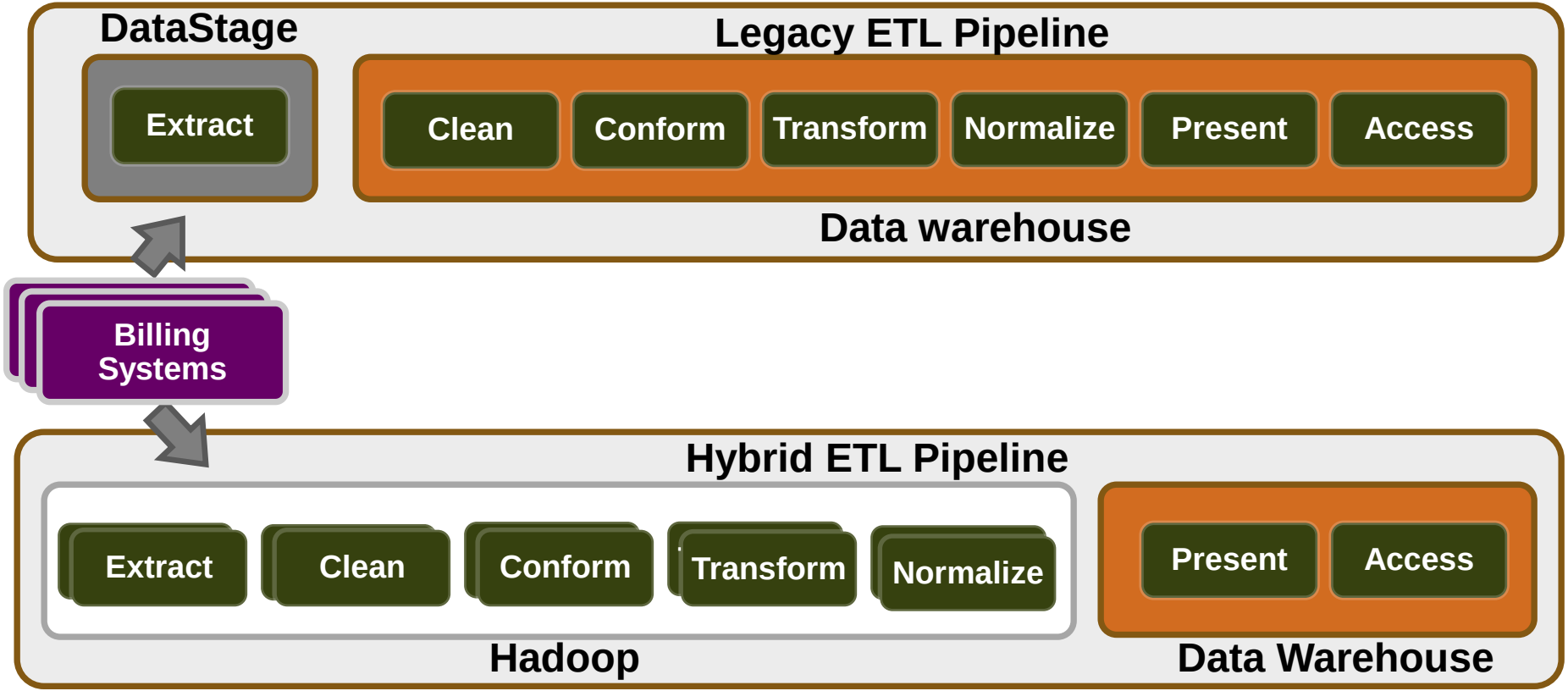


Cost is 1/10th

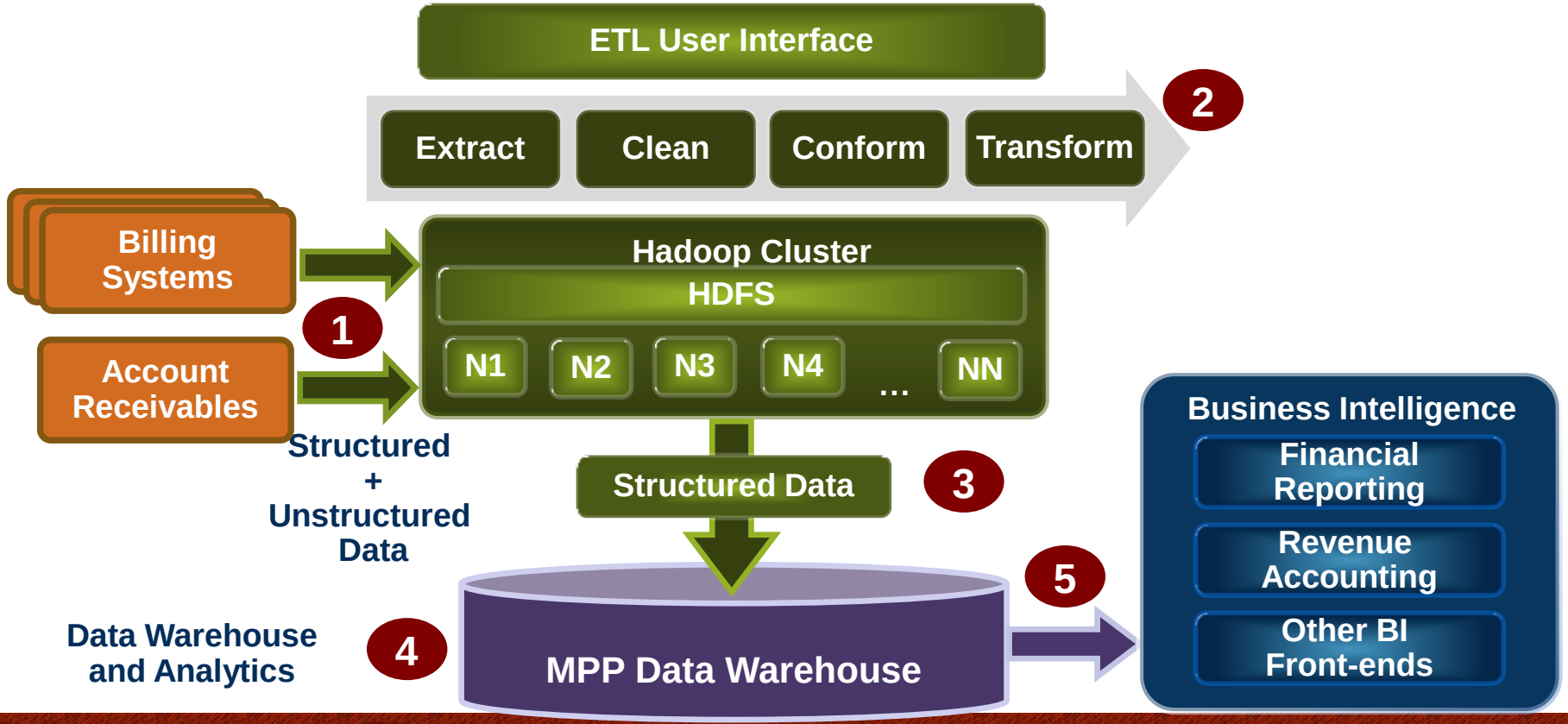
Hadoop Platform Requirements

- Billing, Customer Care, Inventory, Services and other datasets from various sources
- VSAM Flat files (EBCDIC), Message Queue and other sources
- Migration of existing SQL codebase for Transformations
- Production Ready Environment
  - Workflow based Data Pipelines
  - Dependency management
  - Monitoring, Reporting and Logging
- Data Management
  - Metadata based Staging Workflows
  - Policy based Data Retention
  - Data Lineage

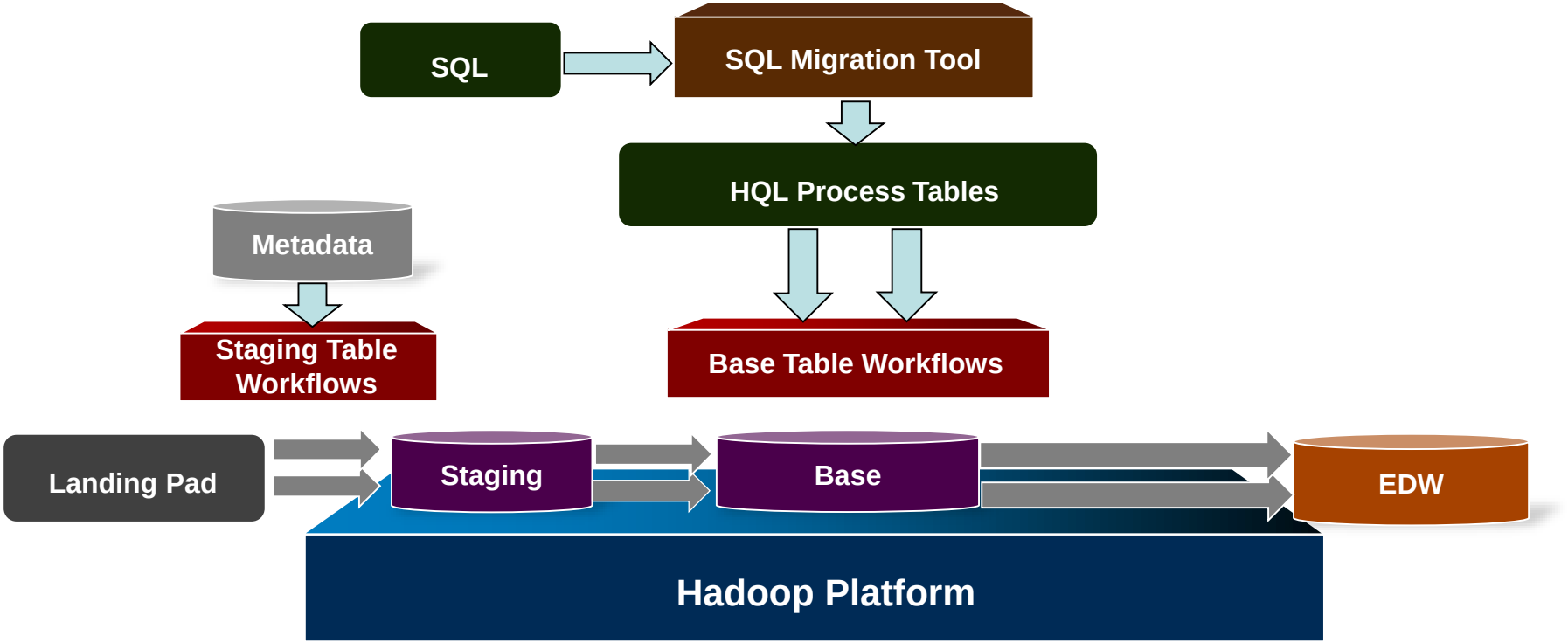
Hadoop ETL Offload



Hadoop ETL Architecture



End to End ETL Solution



EDW Offload – TCO Analysis

Solution	Technology	5-Yr TCO
Existing	EDW	$66,950,000
New	Hybrid: EDW + Hadoop	$33,000,000
	Total Cost Savings	**$33,950,000**

- **CapEx:** Cost avoidance for annual EDW adds
- **Storage:** 20x storage good for next 5 years
- **Cost:** 50x cost reduction (Hadoop = $2,000/TB; EDW = $100,000/TB)
- **Scale Out Architecture:** New nodes can be added on the fly
- **No Disruption:** Hybrid solution requires no changes to upstream/downstream business systems

One Time Hadoop Investment of ~$6.5M Provides $33.9M Cost Savings

Case Study

CAMPAIGN MANAGEMENT

Background

- Customer Insights and ability to target, size, and execute campaigns.

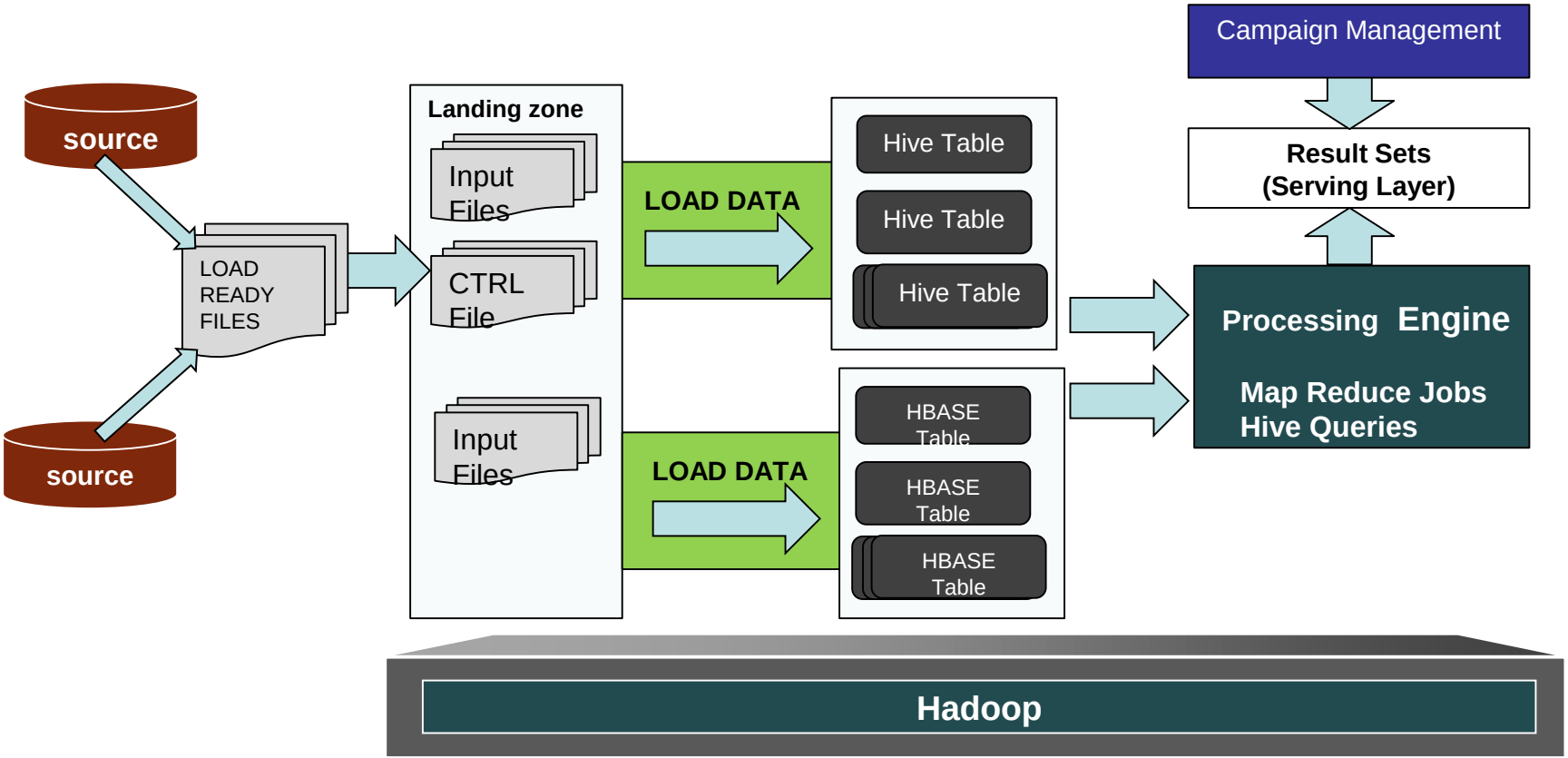
- Intelligent marketing campaigns using Customer behavior based on a Centralized Repository built on the Big data platform.

- Multiple Input Sources
  - Daily Transactional Data
  - User Profiles
  - Rewards and Loyalty Programs
  - Risk
  - Interest Graph
  - Contact History

Campaign Planner Data Model

- Campaign
  - Parent holder for scenario
- Scenario
  - A set of filters to apply to get the result set
- Channel
  - Communication channel like email, SMS
- Module (filters)
  - A set of conditions to apply

Big Data Architecture



Case Study

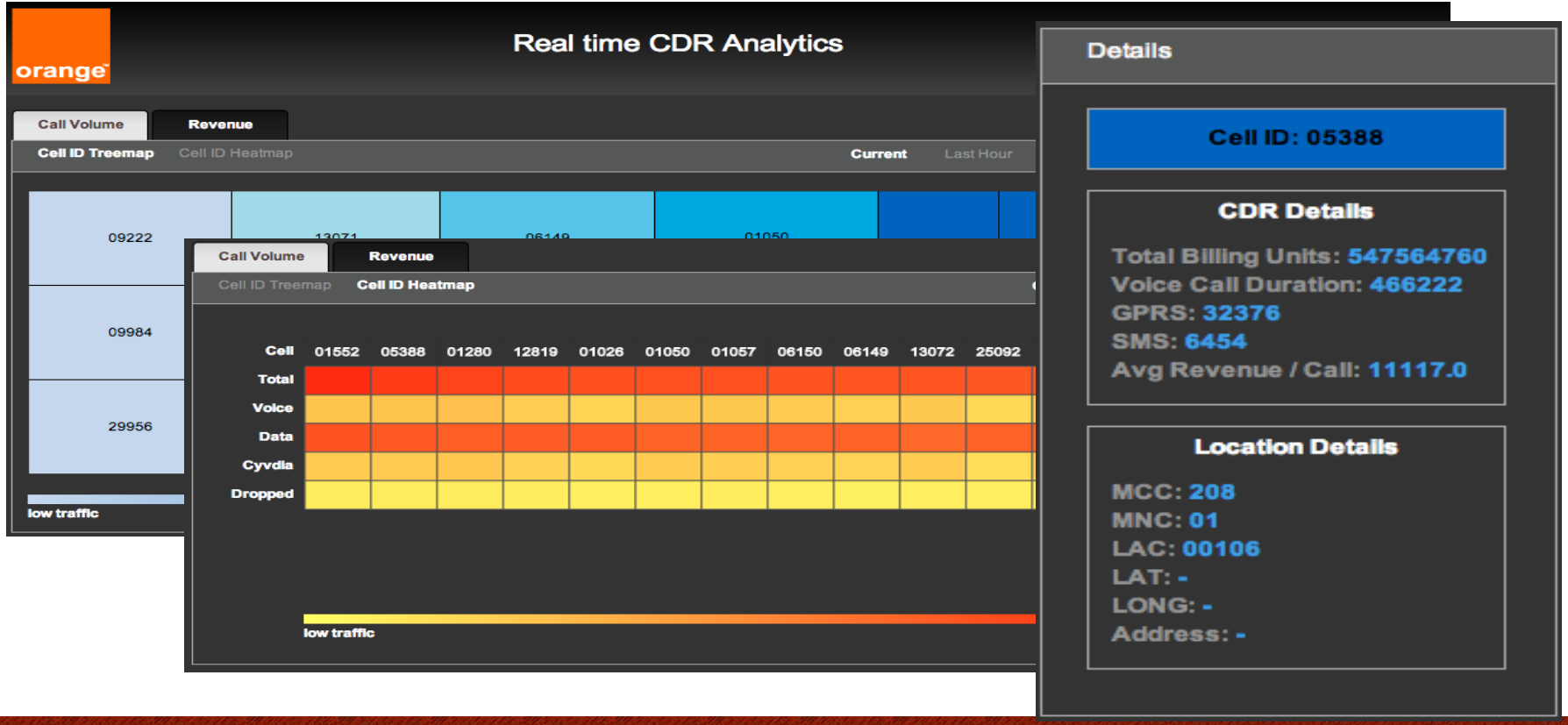
REAL TIME ANALYTICS

Real Time CDR Analytics

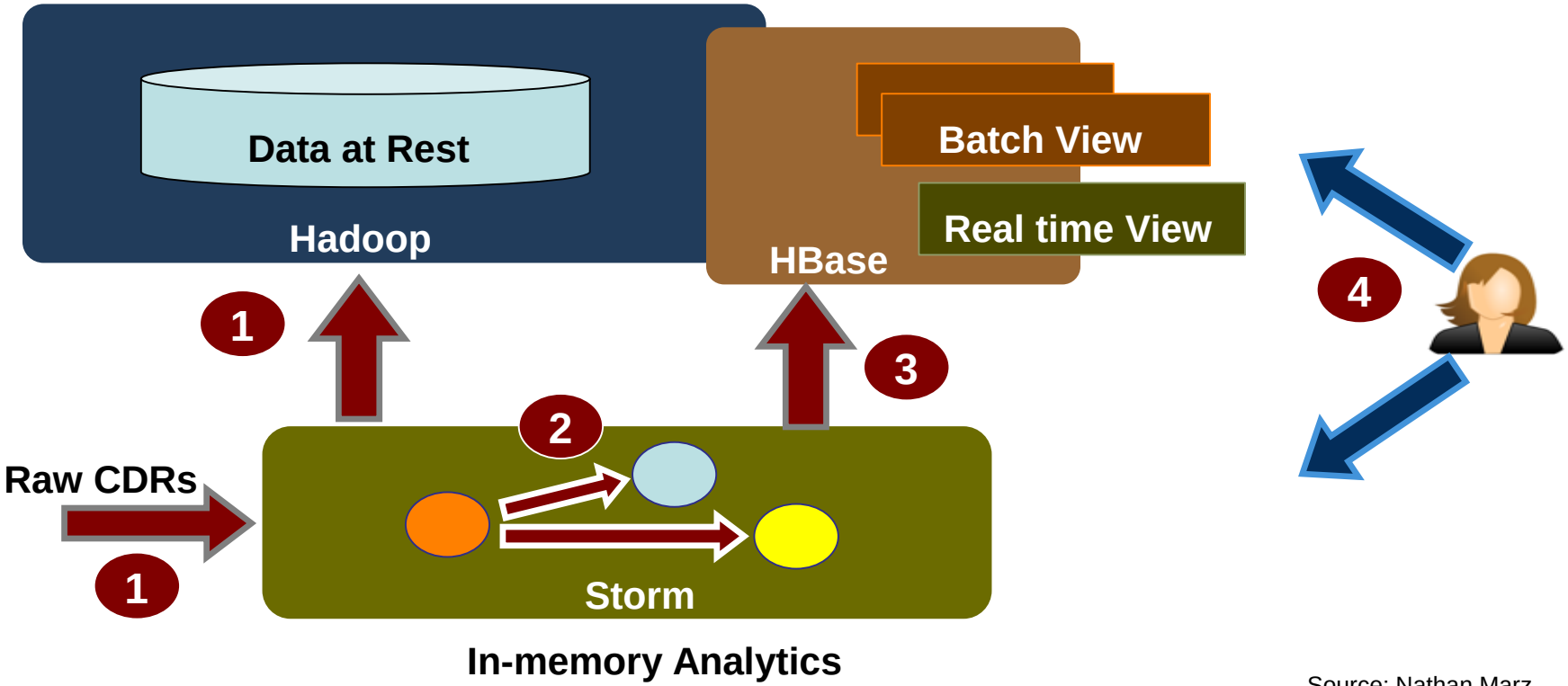
- Real time CDR processing for Network Utilization
  - Network Utilization based on Volume and Revenue
  - Detect Abnormal Termination (Dropped Calls)
  - Filtering for Fraud application

- Machine Learning for Customer Profiling
  - Used Clustering to mine the data and create clusters of customers based on their usage patterns

Visualization

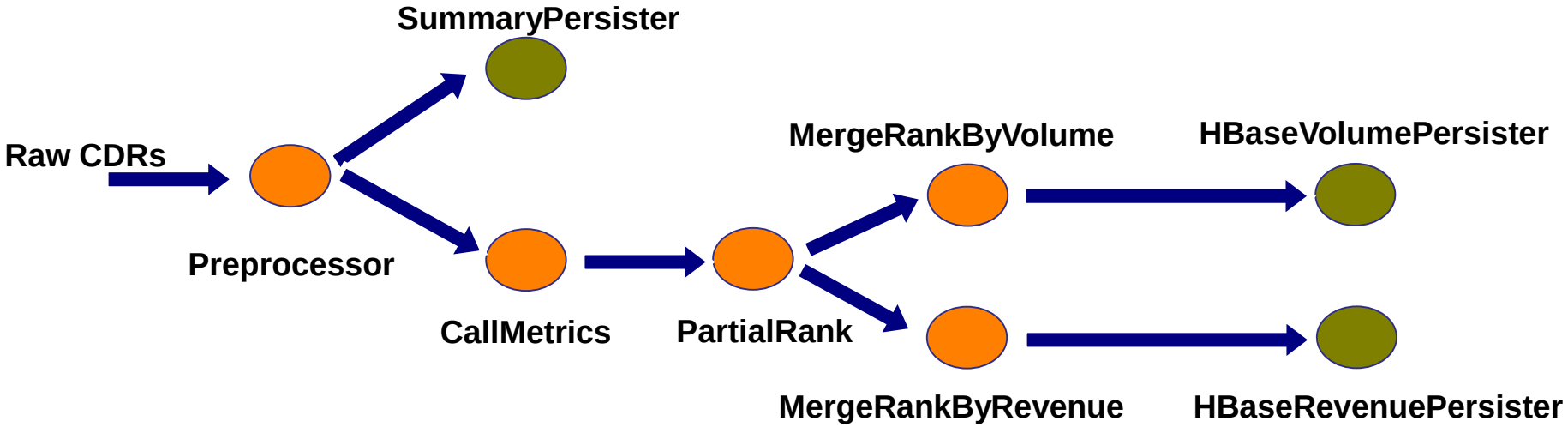


Solution Architecture



Source: Nathan Marz

Stream Processing Topology



Clustering the CDR Data

- Leveraging CDR data to mine customer usage information
  - Segmenting customers based on usage characteristics
    - ✓ Voice call usage
    - ✓ Text messaging usage
    - ✓ Internet usage

- Attempt to overlay our cluster analysis with geolocation data in order to provide tools that can be used for marketing purposes.

- Overlaying with geolocation can tell us about where each cluster congregates

Cluster Characteristics

Cluster-id	n	Calls Made	Ave Call Length	Ave SMS Length	Ave Internet Length	Characteristics
1	129,193	11	703	3	210	Low number of calls but high call length. Very low messaging but high internet. Communicate more through email and voice than text. Affluent people.
2	2,637,226	29	41	11	230	Low on calls and messaging but high on data.
3	1,379,137	42	43	17	444	High on data, people who browse primarily with their phones.
4	3,772,564	53	34	14	23	Older generation; w/o data plans. Note that this is the largest cluster.
5	371,713	448	6	146	90	Exact opposite of cluster 1. Relies mostly on SMS for communication. Even when they get a call, they quickly hang up and switch to messaging. Most socially active people.

Many of these calls could also be SMS and MMS messages

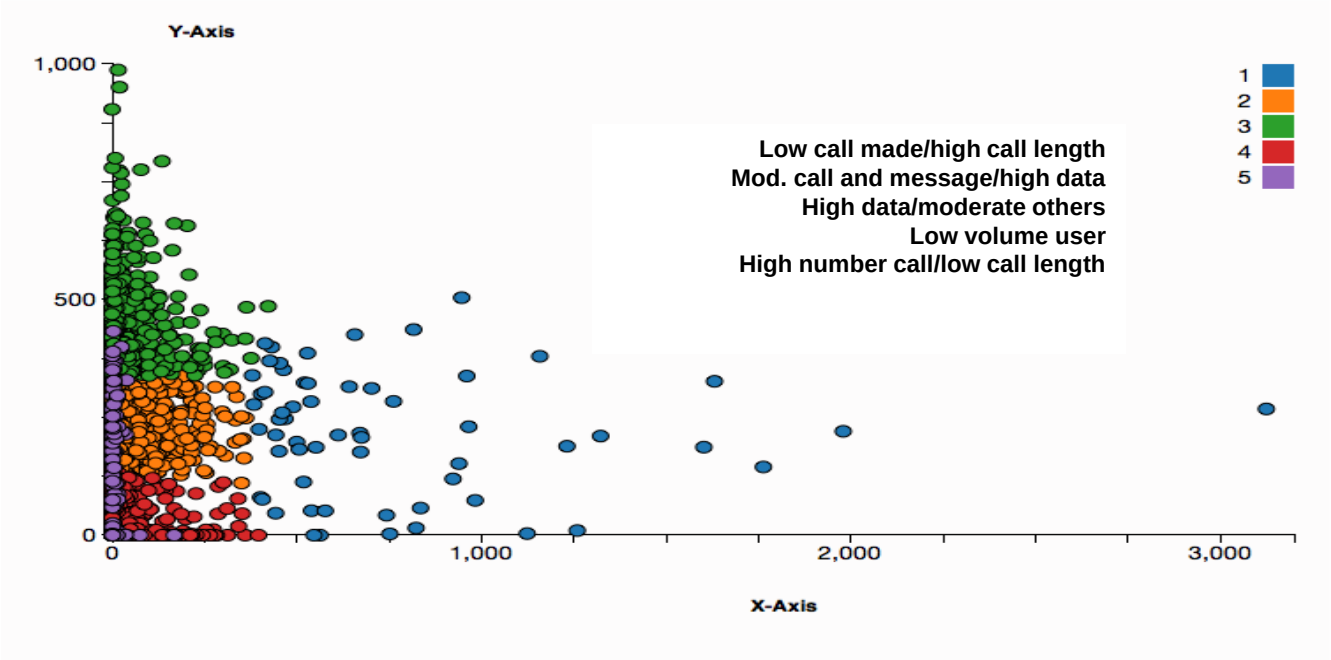
Visualization of the Clustering Data

- Cluster visualization by various dimensions
  - Locality of people by cluster
  - Geolocation by handset usage
  - Geolocation by internet usage
  - Geolocation by phone call usage

- The maps are made at different levels of resolution to show how a marketing person can zoom in and out.

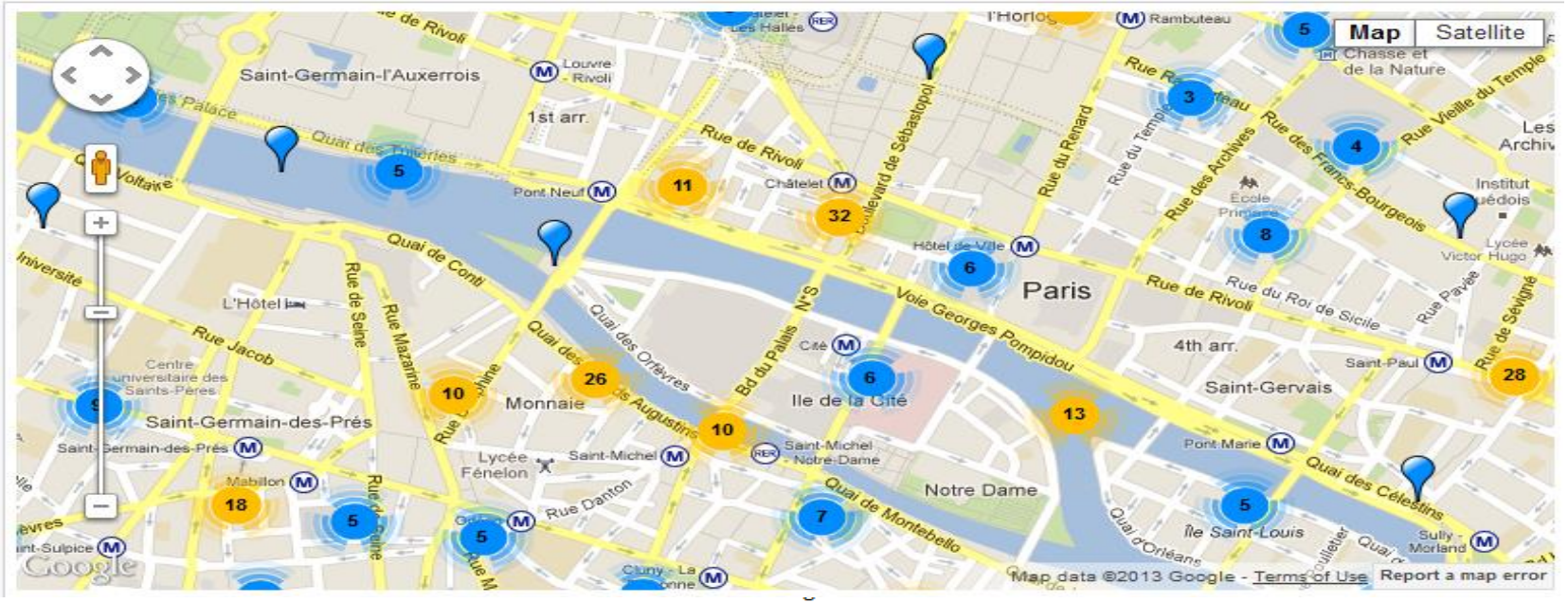
Cluster creation - Calls Versus Data

Plot of total calls made versus internet data usage. Gives good separation between red, orange, and green clusters (lowest to highest internet users).



Geolocation at the Cluster Level

Here we plot the locations of calls made by people in Cluster 4 zoomed in even more.



ZALONI

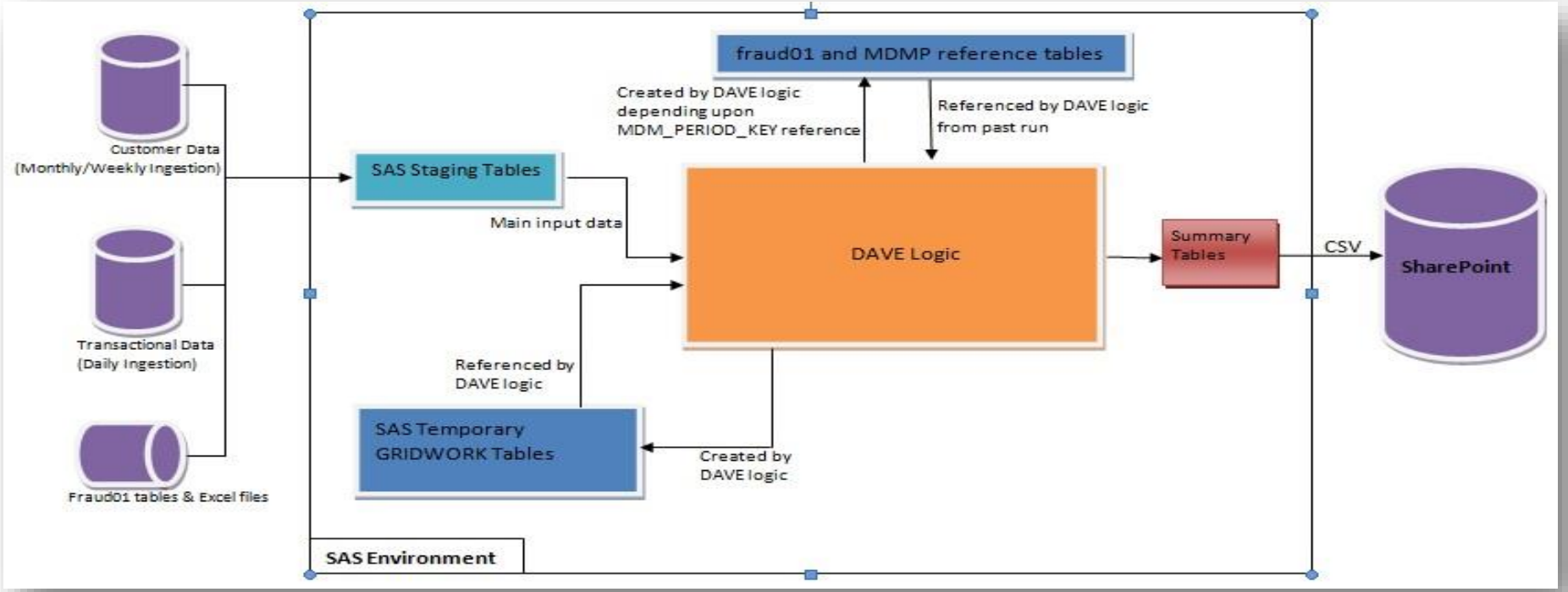
Case Study

Check Fraud Detection in Hadoop

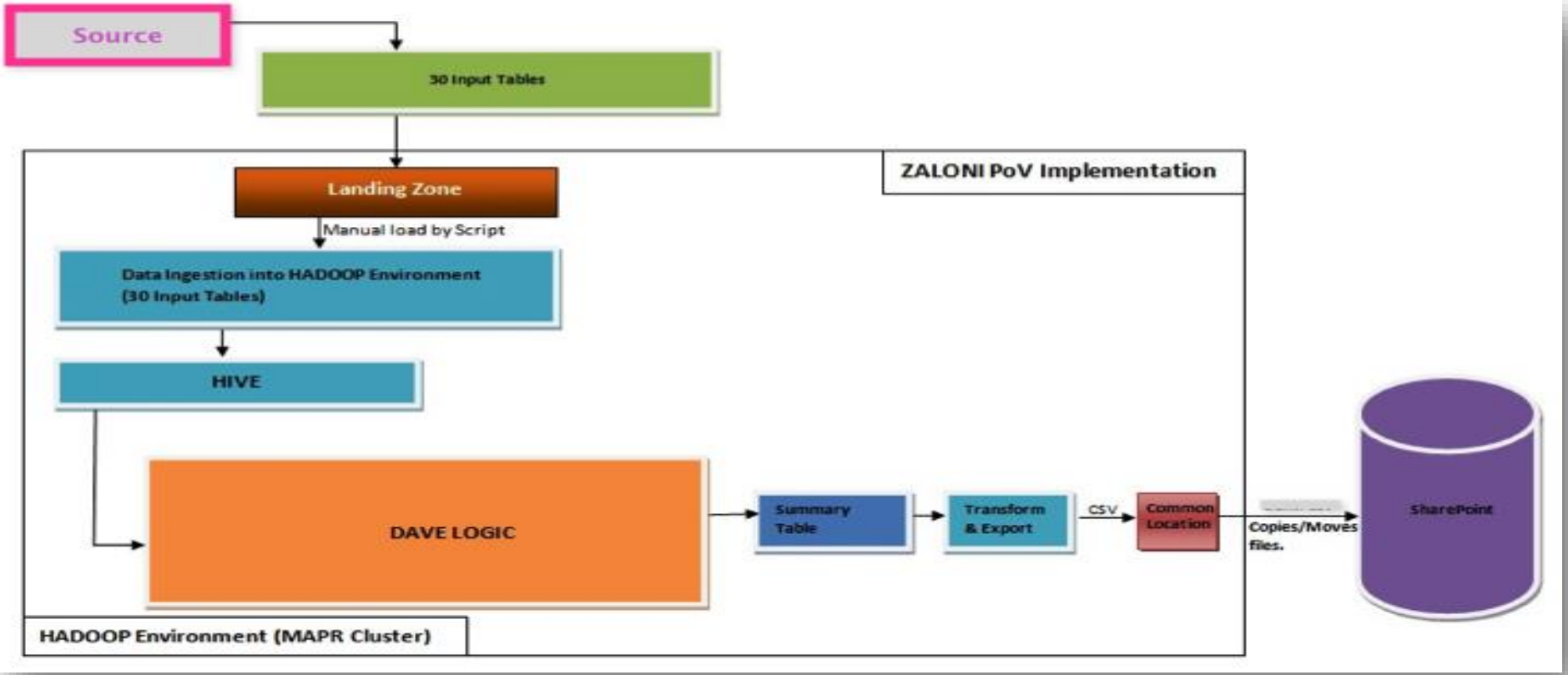
Background

- Major US Bank performs ATM check deposit fraud analytics on a daily basis
- Use SAS-based analytics system referred to as DAVE which contains the primary business logic that performs daily fraud analytics and generates alerts and reports of potential ATM check deposit fraud transactions.
- Source ATM transaction data comes from upstream mainframe banking systems. In addition, DAVE queries other structured reference data from Enterprise Data Warehouse (EDW) systems
- Daily fraud alerts generated by DAVE are posted on an internal SharePoint environment for further analysis by the Enterprise Fraud Analytics group.

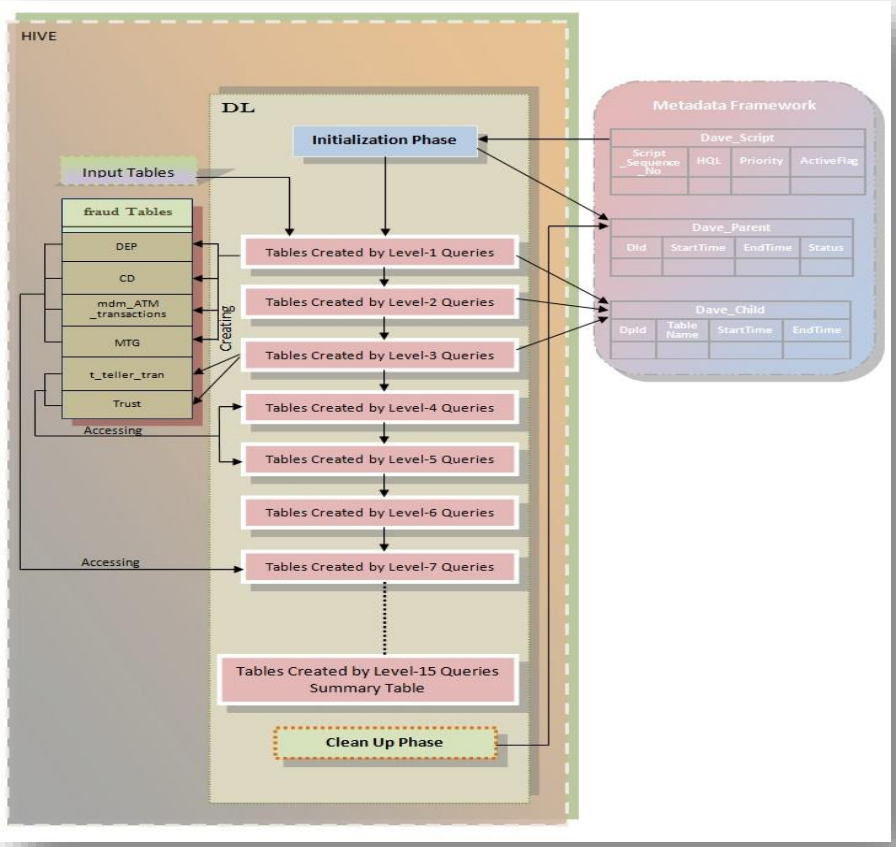
Existing Architecture



Hadoop Solution Architecture



Hadoop Solution Architecture(contd..)



- Dave logic implementation in Hive highlights
- Queries are picked up from a metadata framework and executed in priority order
- Temporary tables are variables may be created, are cleaned up in the end
- Queries having same priority are executed in parallel.
- Execution model written in Java

Data Volume

- Data Sources
  - DB2
  - Oracle
  - Excel
- Volume of Events per day
  - It was a PoV so client had given 2 months of data
  - There are 30 input tables, each table has an average of 10-12 GB of data.
- Input data was flat & delimited files

Sample Query for Summary Table Creation

```
CREATE TABLE fraud02_Alerts_Filtered_Sysdate_User_Name AS select *, CASE
WHEN  Category_ = 'Hot ATM: Known Fraud Ring' THEN 1.0 WHEN  Category_ = 'Hot ATM: ANR Hard Alert' THEN 2.0
WHEN  Category_ = 'Hot ATM: ST-to-ST Hard Alert' THEN 3.0 WHEN  Category_ = 'Hot ATM: Starter-checks'  THEN 8.0
WHEN  Category_ = 'Mobile: ANR Hard Alert' THEN 4.4 WHEN  Category_ = 'Mobile: ST-to-ST Hard Alert' THEN 4.5
WHEN category IN (""23_BAD_ANI"") THEN 4.55 WHEN  Category_ = 'Mobile: ANR Hard Alert' AND total_deposit_day >= 800 THEN 4.0
WHEN  Category_ = 'Hot ATM: Minimum history' THEN 4.6 WHEN category = ""23_High_Risk_Mobile"" THEN 4.75
WHEN category = ""23_High_Risk_NAC"" THEN 4.76 WHEN  Category_ = 'Mobile: Minimum history' THEN 4.8
WHEN New_Rules != ""Else "" THEN 5.0 WHEN  Category_ = 'Hot ATM: Dep > Avg' THEN 5.5
WHEN  Category_ = 'Hot ATM: NAC Account, ATM deposit' THEN 6.0 WHEN  Category_ = 'Hot ATM: No ATM History'  THEN 7.0
WHEN  Category_ = 'Hot ATM: Largest deposit in 4 months' THEN 9.0 ELSE 99.0 END AS Ranking from ( select *,CASE
WHEN ATM = 0 THEN concat(""Mobile: "",new_category) ELSE concat(""Hot ATM: "",new_category) END AS Category_
from ( select DISTINCT ATM,st_deposit_acct,total_deposit_day,category,New_Rules,CASE WHEN FRDABA IN
('103104900','061119794','103101673','102100400','091215558','091900533') THEN ""Known Fraud Ring""
WHEN category IN (""00_ICMS_Alert"",""21_Known_Bad_Check"",""00_EWS_Alert"") THEN ""ANR Hard Alert""
WHEN category IN (""23_BAD_ANI"") THEN ""ANR Hard Alert2"" WHEN category = ""00_Drawn_on_CLOSED_Account"" THEN ""ST-to-ST Hard Alert""
WHEN category = ""23_High_Risk_Mobile"" THEN ""Dep > Avg"" WHEN category IN (""04_Suspected_Kiter"") THEN ""Dep > Avg""
WHEN category IN (""04_1+ Years extreme factor_A"",""04_1+ Years extreme factor_B"",""04_1+ Years extreme factor_C"") THEN ""Dep > Avg""
WHEN category = ""04_1+ Years first ATM and $4000+"" THEN ""No ATM History, large deposit""
WHEN category = ""15_No_ATM_Deposits_in_4_Months"" THEN ""No ATM History"" WHEN category = ""14_Period_1_Account"" THEN ""Minimum history""
WHEN category IN (""16_Total_Deposit_GT_MAX_A"",""16_Total_Deposit_GT_MAX_B"",""16_Total_Deposit_GT_MAX_C"",""16_Total_Deposit_GT_MAX_D"") THEN
""Largest deposit in 4 months"" WHEN category = ""18_NAC_Account"" THEN ""NAC Account, ATM deposit""
WHEN category = ""17_Deposit_GT_ATM_Average+STDEV"" THEN ""Dep > Avg"" WHEN category = ""19_Suspicious_Routing_Number"" THEN ""Known Fraud Ring""
WHEN category = ""20_Another_Check_On_Hold"" THEN ""Dep > Avg"" WHEN category = ""22_Deposited_starter_checks"" THEN ""Starter-checks""
WHEN category = ""23_High_Risk_NAC"" THEN ""Dep > Avg"" END AS new_Category, OTHR_HOLD_AMT_Total, CITY_NM, ST_CD, atmid,
CASE WHEN category = ""15_No_ATM_Deposits_in_4_Months"" AND ATM != 1 THEN ""DELETE"" ELSE ""Keep""
END AS Delete_Flag, CASE WHEN Total_Deposit_Day > 10000 THEN 10000 ELSE Total_Deposit_Day END AS Total_Deposit_Day_ceiling
FROM fraud01_HOT_ATM_Alerts3_Sysdate_User_Name WHERE selected = 1 ) t1 where t1.Delete_Flag != ""DELETE"" ) t2"
```

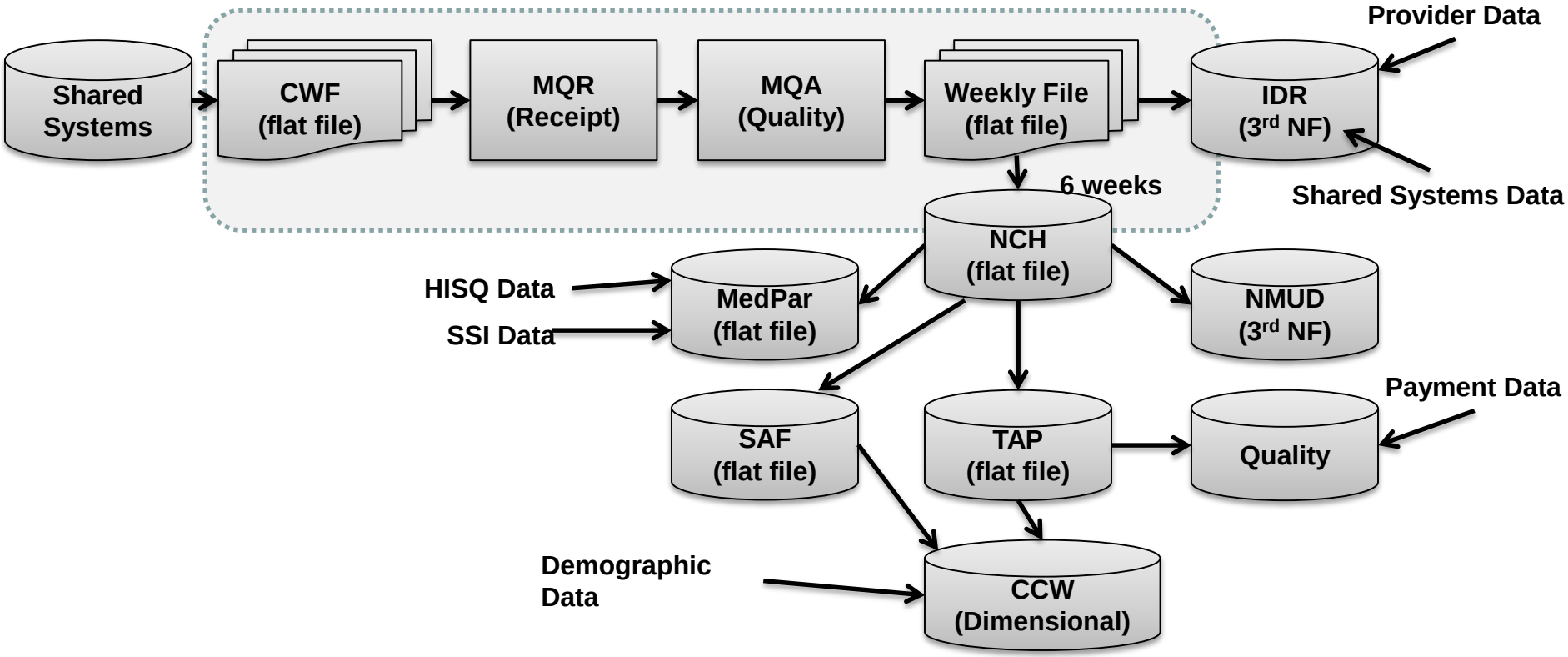
Case Study

MEDICARE KNOWLEDGE DISCOVERY INITIATIVE

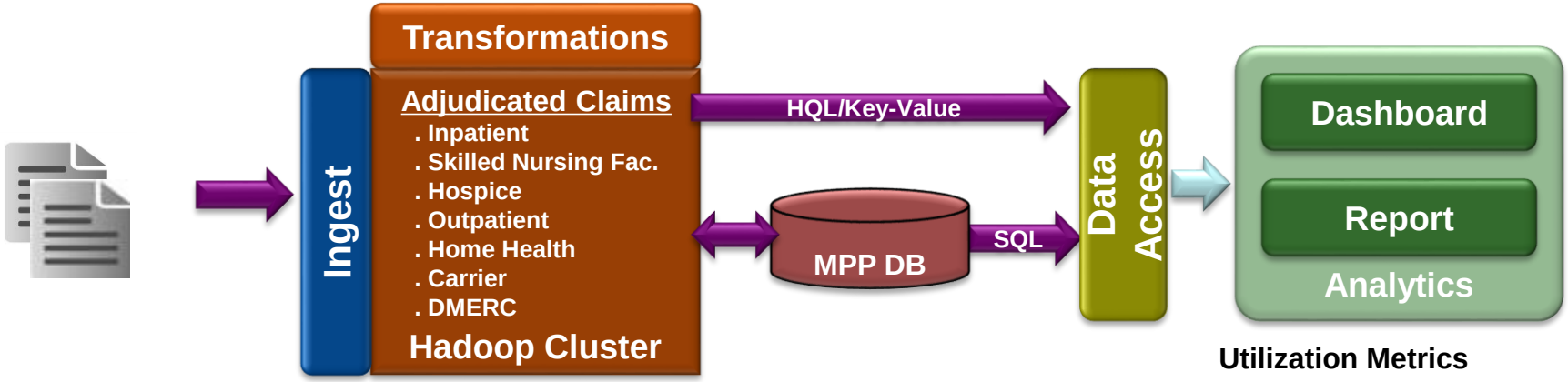
Value Proposition

- Reduce time to market CMS claims data for analytics (currently 6 weeks)
- Reduce risk with legacy mainframe programs including MQR, MQA and NCH (complex logic with fewer available SMEs)
- Increase and optimize IDR capacity by reducing or eliminating ETL in IDR (free-up of ~38TB)
- Reduce data duplication (MQA, NCH, CCW, IDR,,, many more)
- Enable new analytics possibilities in support of CPI, CMMI, etc

Claims Processing



High Level Solution



Adjudicated Claims (Daily)

- Inpatient
- Skilled Nursing Facility
- Hospice
- Outpatient
- Home Health
- Carrier
- DMERC

Cummulative/Monthly/Weekly

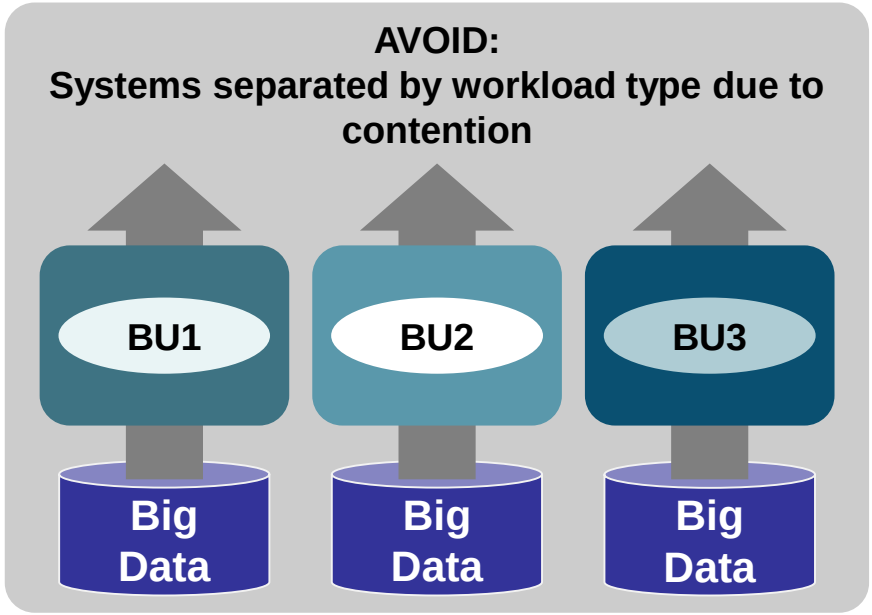
SRVC_YR=2012

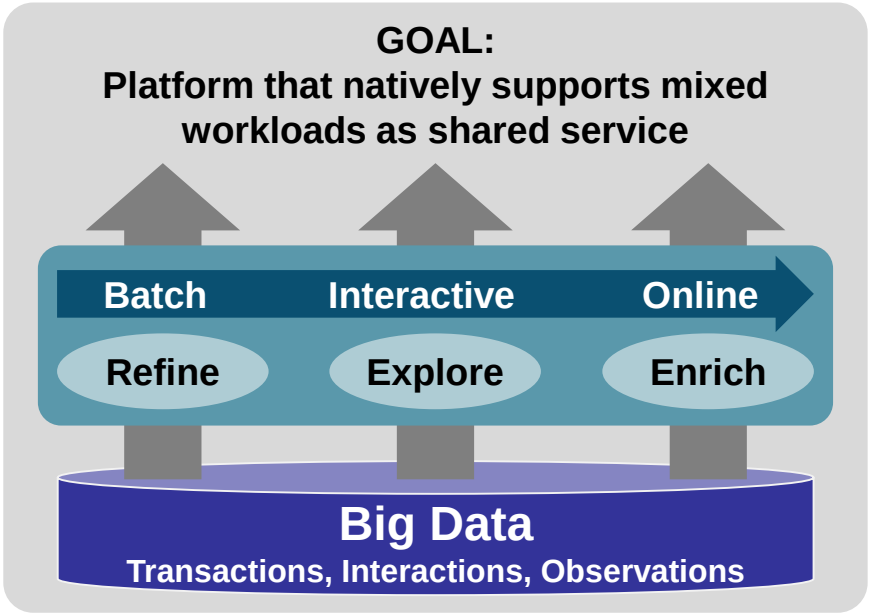
OBS	TYPE	GROSSCLM	NETCLAIM	NETREIMB
148	CARR	685,805,053	670,306,475	68,221,272,748.23
149	DME	58,005,869	55,144,353	7,305,119,254.70
150	HHA	15,418,485	5,948,433	14,514,559,367.68
151	HSP	3,313,132	3,065,366	10,641,698,376.33
152	INP	13,161,878	11,563,730	98,989,747,976.76
153	OUT	127,670,903	119,621,765	40,732,560,311.51
154	SNF	4,543,448	4,101,418	20,200,288,452.18
SRVC_YR		907,918,768	869,751,540	260,605,246,487.39
		21,508,241,616	19,767,144,714	4,825,921,786,022.08

Strategy for a successful

DATA LAKE IMPLEMENTATION

Strategy: Data Silos to Data Lake





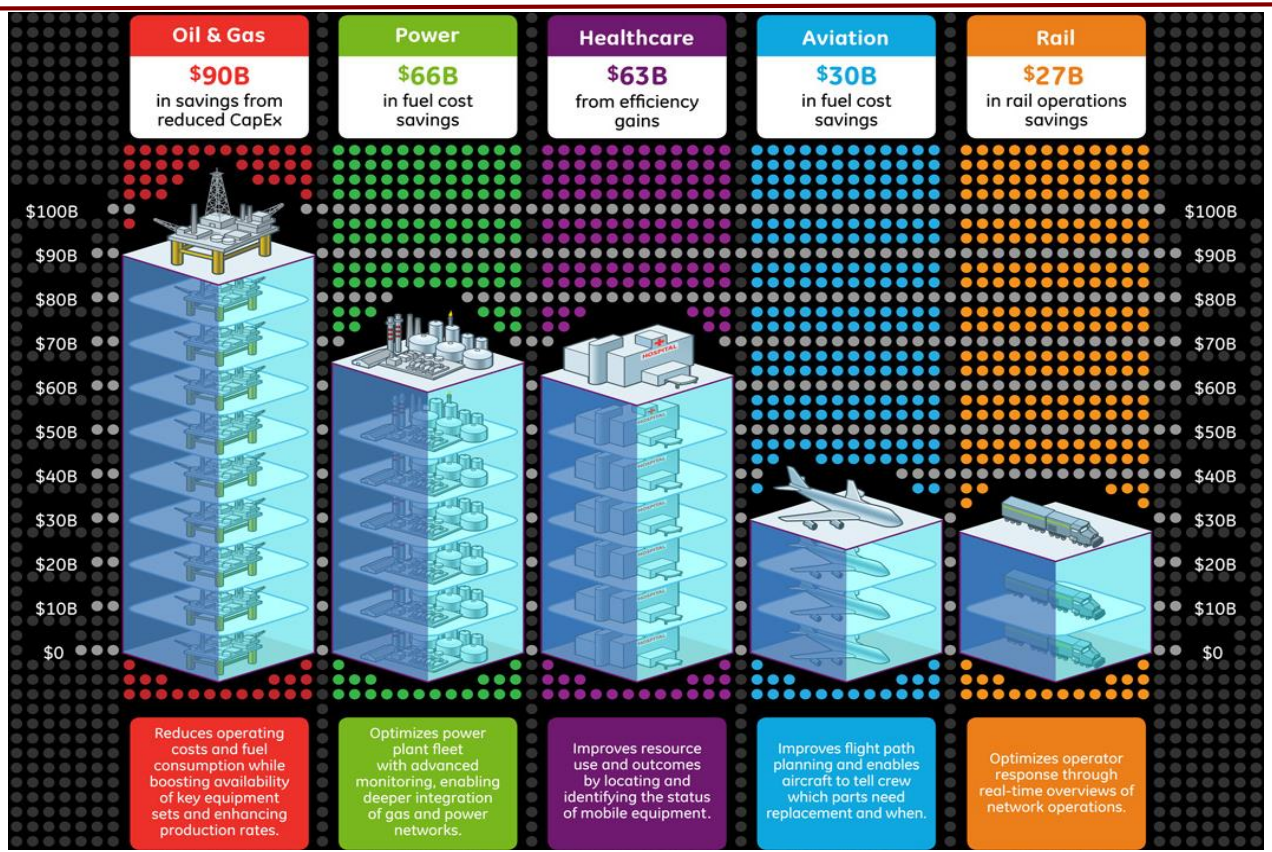
Challenges

- Identity the big data problem
  - Add a validation criteria
  - “Is this really a big data problem”
- Development skillset
- Integration with current and legacy systems
- Security and Compliance
- Augmenting existing operations team to manage production data lake

Proof of Technology for Data Lake

- Demonstrate business value
- Time bound
- Specific success criteria

GE – Potential value of investment



Source: GE Industrial Internet Pushing the Boundaries of Minds and Machines

Proof of Technology considerations

- POT Objectives
- Infrastructure considerations
- Timeline and Resources
- POT Readout

Data Lake Center of Excellence

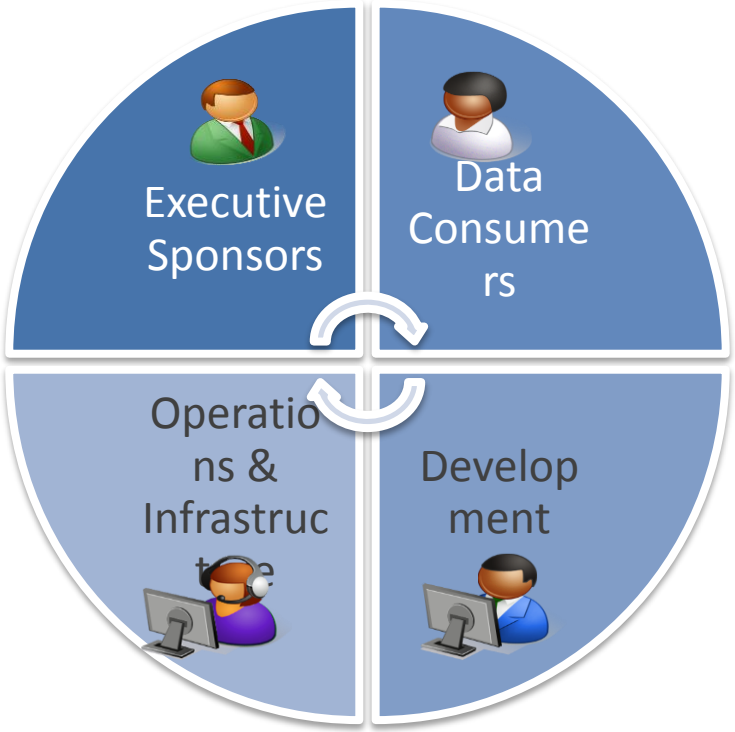
Why build one?

- Unified Data Lake instead of silos
- Common and consistent process for ingestion and organization and extraction of data
- Unifiy talented pool of resources

Data Lake Center of Excellence

ZALONI

Key Roles
- Executive Sponsor
- Program Director
- Chief Architect



Key Roles
- Data Scientist
- Business Unit Owner
- Data Steward

Key Roles
- Operations Manager
- Infosec Manager
- Infrastructure Architect

Key Roles
- Development Manager
- Architect
- IT Developers
- Data Governance

Architecture characteristics

- Leverage frameworks
- Create reusable components
- Always test for scale
- Support agile development
- Data Audits and Lineage
- Security

Backup

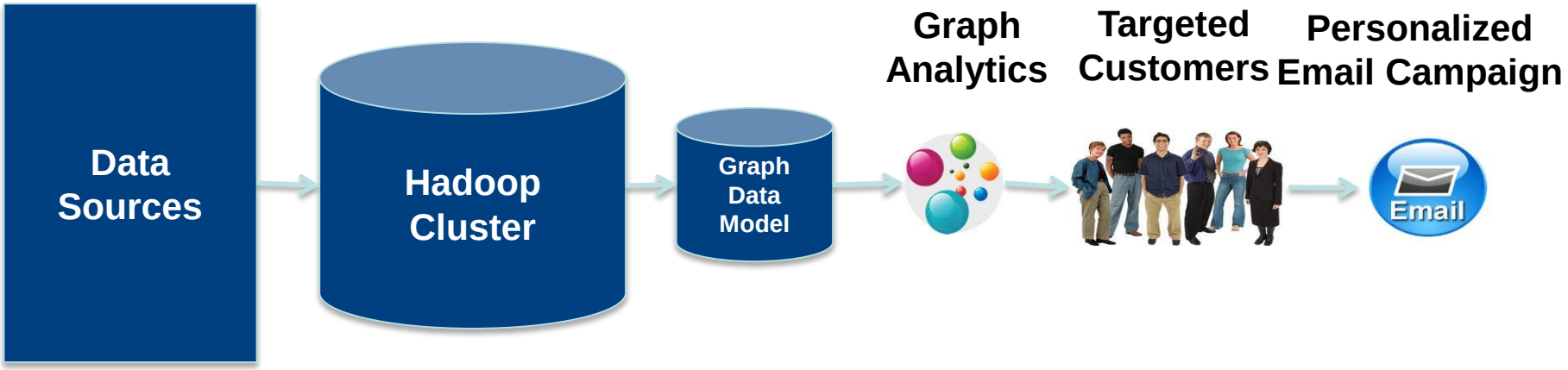
Case Study

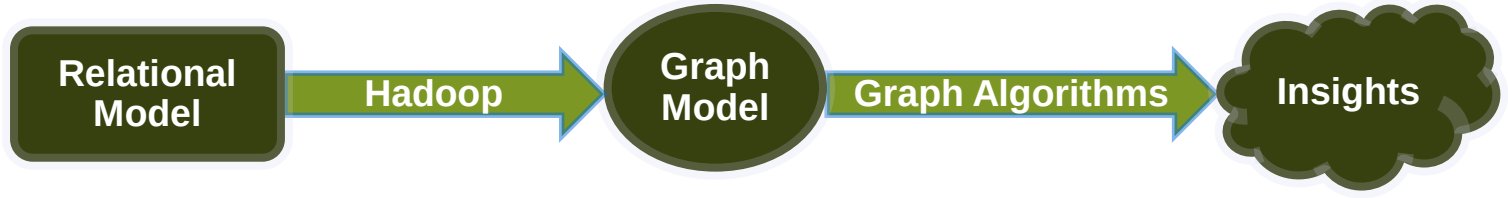
PERSONALIZED MARKETING USING HADOOP AND GRAPH DB

Overview

- Analyze millions of transactions from Brick & Mortar and Online stores
- Capture as many data sources as possible into Hadoop
- Build graph data model using Hadoop
- Execute graph analytics to drive new “email personalization” campaign
- Significantly improve email campaign conversion/lift.
- Establish case to deploy “web personalization” next.

Use Case

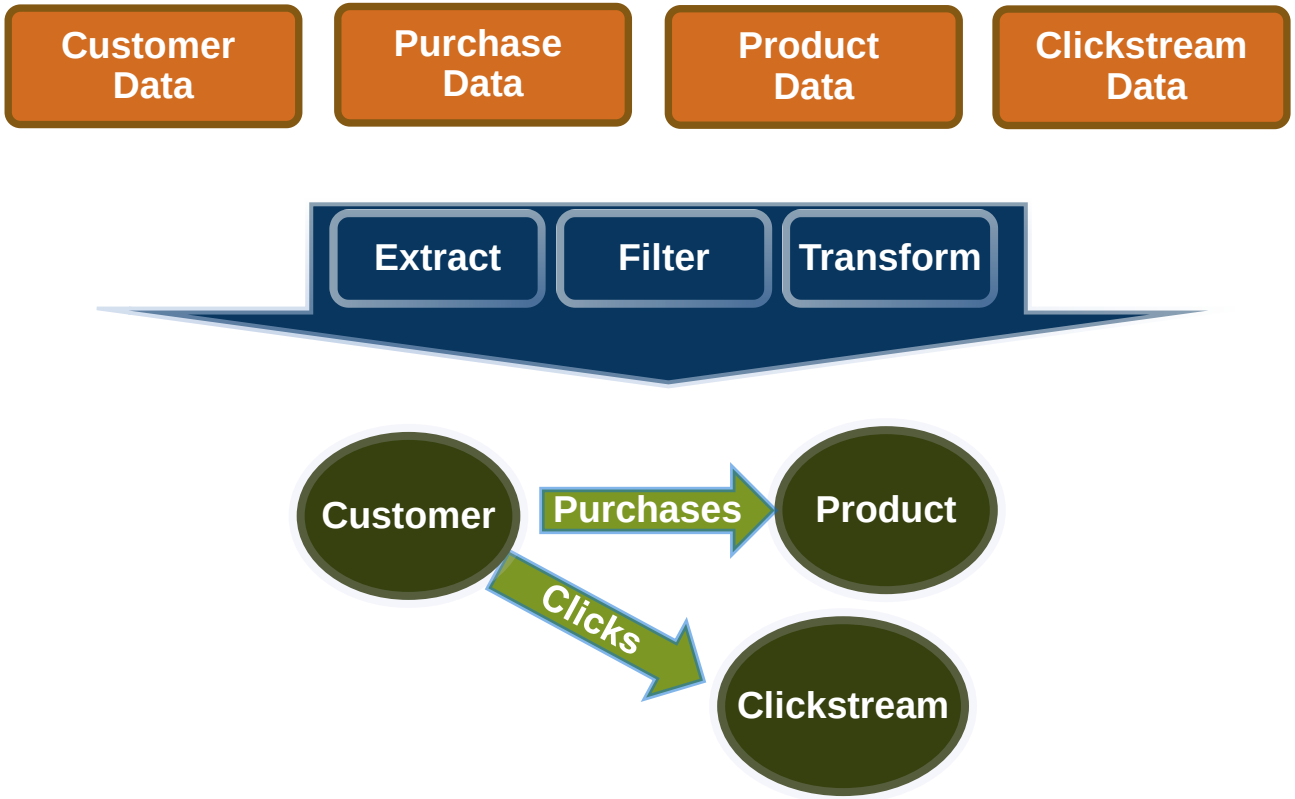




Quadstores, Attributes and Graphs

GRAPH MODEL CREATION

Graph Model Creation



Storing the Graph Model: Quadstore

- The directed graph contains millions of Customer, Product, Clickstream vertices, and Clicks and Purchases edges.
- The graph is stores as a quadstore file, which correlates edges with vertices:

Subject Vertex ID	Edge Type	Edge ID	Object Vertex ID
1	1	1	4
1	0	2	3
2	1	3	6
2	0	4	4