





# MODERN DATA WAREHOUSE FUNDAMENTALS

Part I: Introducing the Modern Data Warehouse - Challenges, Use Cases, and Opportunities

December, 2018

# SPEAKERS



**Eva Nahari**

Director, Product  
Management  
eva.nahari@cloudera.com



**David Dichmann**

Director, Product Marketing  
ddichmann@cloudera.com

**cloudera**

---

# Why Modernize Your Data Warehouse?

The Case for a Modern Data Warehouse

# LARGE NORTH AMERICAN BANK

## FRAUD PREVENTION

247

Users

2M

Queries / Month

500

Terabytes

167

Databases



- LoB Data Analysts **access all data**
- Saved **\$4M+** in deposit fraud



# GLOBAL PHARMACEUTICAL

## NEW PRODUCT DEVELOPMENT

8700

Users

8 PB

Diverse Data

200+

Use Cases

200x

Fewer Silos



- Curated Use and Agile Discovery with **HIPAA compliance**
- Accelerated new **Drug Development**

# MAJOR TELCO MANUFACTURER

## BUSINESS OPTIMIZATION

7

New Sources

2M

Users

<1S

Query  
Responses

>100 M

Min. Data Sets



- **\$10 M new revenue** from optimized marketing
- **\$30 M+** from Price Optimization
- **\$100K+** from weather correlation

# NEW TRENDS IN DATA WAREHOUSING

Deeper Business Insights at Extreme Speed and Scale While Managing Cost





# NEW TRENDS IN DATA WAREHOUSING

## Deeper Business Insights

### Protect

- Proactive Fraud Prevention
- Keep up with Regulatory Compliance
- Preempt Cyberthreats

**Real-time response on massive data volume and variety**

### Grow

- Customer Sentiment
- Fault Prevention
- Improve Product Quality
- New Revenue Streams

**Experimentation and collaboration at scale**

### Optimize

- Improve Operational Efficiency
- Support Internet of Things (IoT)

**New analytics techniques democratized to all users**

# NEW TRENDS IN DATA WAREHOUSING

## Extreme Speed and Scale

### More Data

- Massive amounts handled faster at scale
- More variety from new sources (social media, IoT)
- Insight within minutes of new data arrival

### Performance and flexibility at scale

### More People

- 1,000's of new users and new user types
- 1,000's of new use cases
- All skill levels: Analytics, Data Science, and Machine Learning

### All workloads with a shared data experience

### More Workloads

- 100's of production grade deployments
- Enterprise grade dependability
- Strict security and governance

### On-demand scale out, discovery, collaboration

# NEW TRENDS IN DATA WAREHOUSING

## Managing Resources and Costs

### Optimize Core Processes

- Automation to reduce pressure on organizational bottlenecks
- Consistent user experience

**Broaden data reach without increasing IT burden or costs**

### Dynamic Consumption

- Transient Workloads
- Short-lived Workloads
- Permanent Workloads
- Public, Private, Hybrid Cloud

**Environmental flexibility and adaptive compute, storage**

### Self-Service Everything

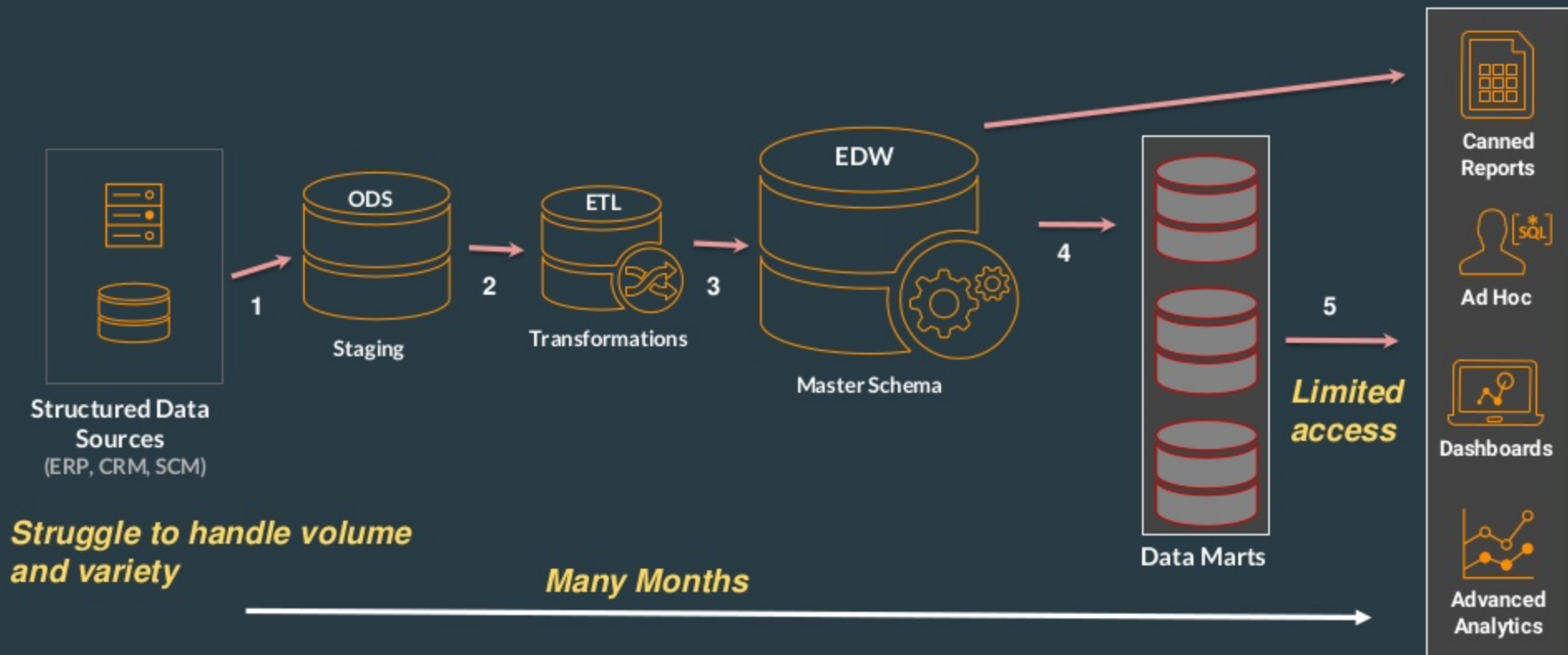
- Resource provisioning
- Workload development
- Optimizing and troubleshooting

**Deliver on increased SLA pressures without runaway cost**

**Quickly** enable business analytics by sharing **petabytes** of **verified** data across **thousands** of users while **surpassing** demands of **SLAs** and costs



# TRADITIONAL DATA WAREHOUSE:



# WHAT CONCEPTS SURVIVE?



**Data Modeling**



**Security & Governance**



**Reports & Dashboards**

# WHAT HAS CHANGED?

## Traditional DW

## Modern DW

Data & Analytics



Supporting Role

Foundational Role

Users



Primarily Internal

Internal & External

Data Curation



Planned ETLs

On-Demand Pipelines

Data Exploration



Constrained, Structured

Freeform,  
Multi-Structured

# WHAT IS NEW?



**Experimentation  
& Collaboration**



**Dynamic Consumption**



**Self Service  
Everything**

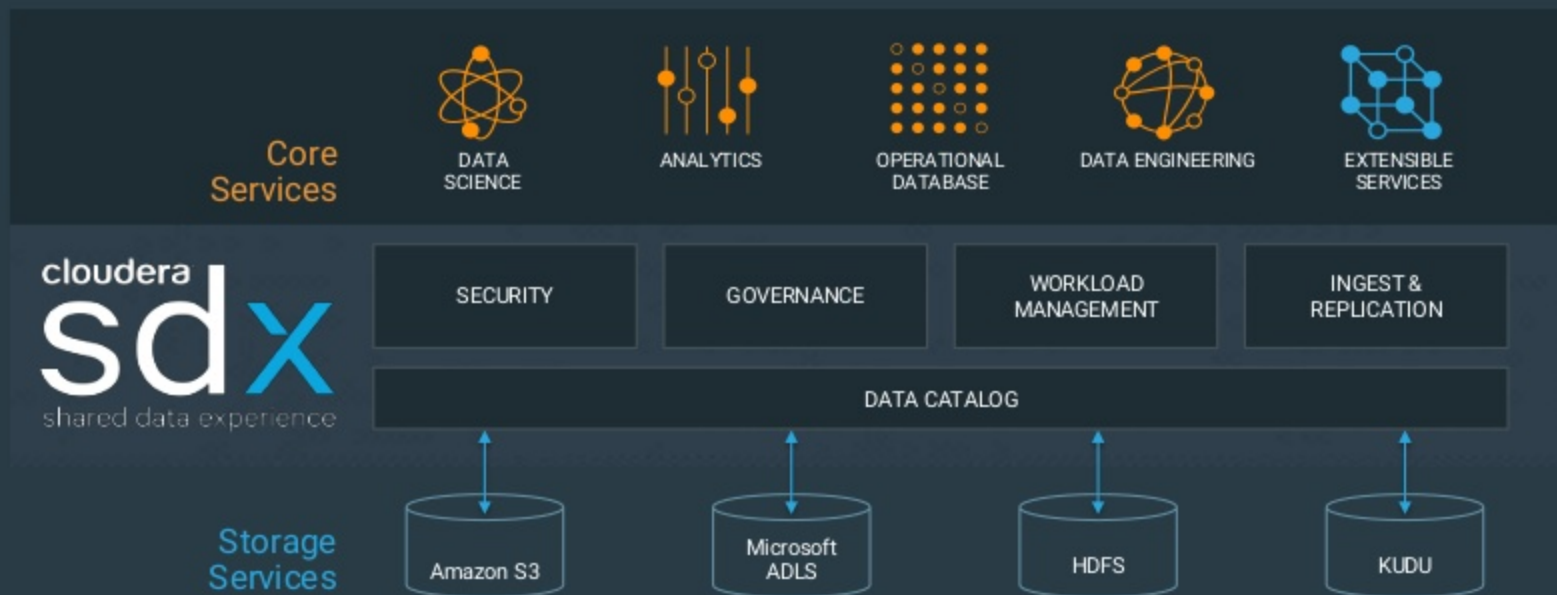


# MODERN DATA WAREHOUSE



# CLOUDERA MODERN DATA WAREHOUSE

The modern **platform** for machine learning and analytics optimized for the cloud



# A MODERN DATA WAREHOUSE SOLUTION

SQL Workbench



Preferred BI & ELT Tools



Workload  
XM



Cloudera  
Manager



Altus



Navigator  
& Sentry



Hive-on-Spark / Spark  
MPP ELT Processing



Impala  
MPP Query Engine



Solr  
MPP Search Analytics

Shared Data Experience (SDX)

Optimized File Formats (Parquet, Avro)

KUDU | HDFS  
Local Storage



AWS S3 | ADLS  
Object Storage

Move faster  
Serve more users  
Reduce IT pressure

**Self Serve** Diagnostics and Optimizations



**WORKLOAD XM**

**Proactively Optimize Workloads**



**Self Serve** Analytics Workbench



# EXTREME SPEED & SCALE

Fastest ELT  
at Scale  
**for Data Engineers**



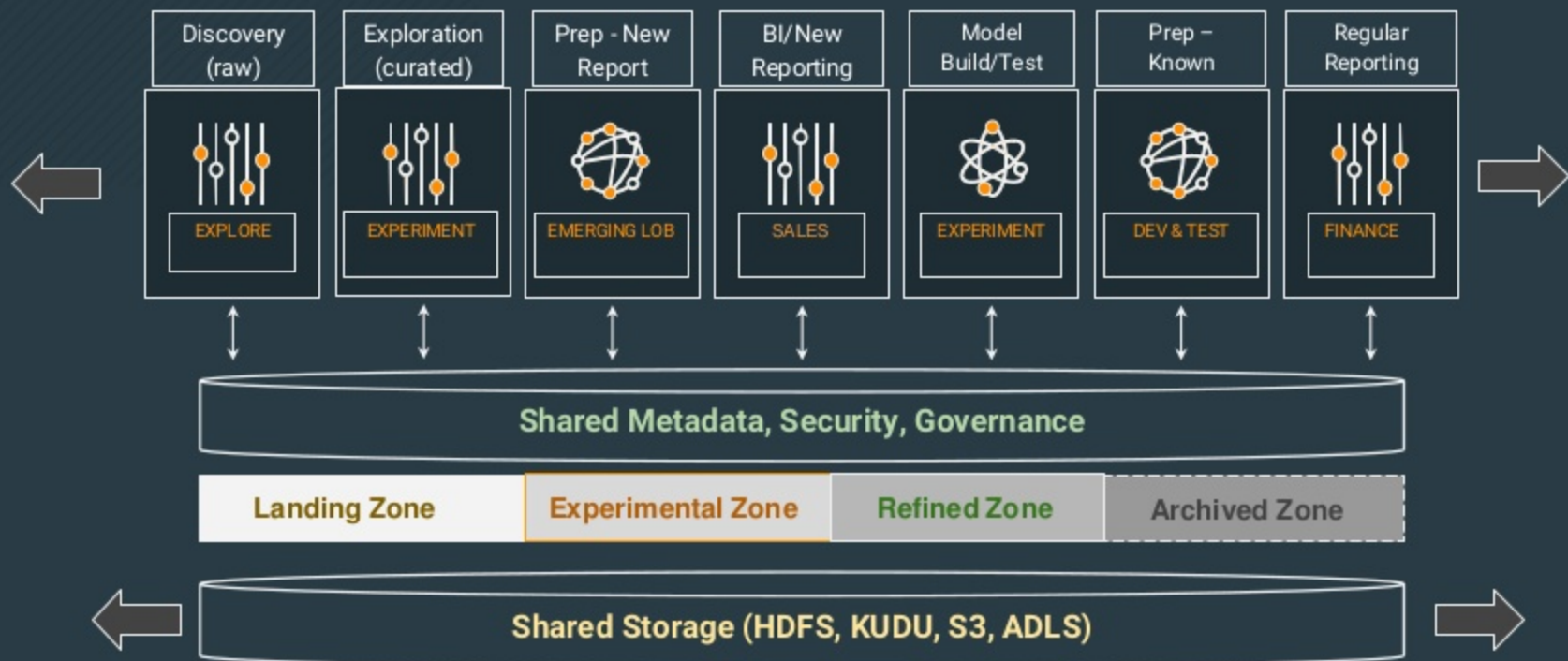
Fastest Self-Service BI  
at Scale  
**for Analysts & Developers**



Flexibility at scale  
1000s of users  
On-demand scale out  
Speed to insight



# ON-DEMAND SCALING & MULTI-TENANCY



# ENABLES FULL FLEXIBILITY AND DYNAMIC CONSUMPTION

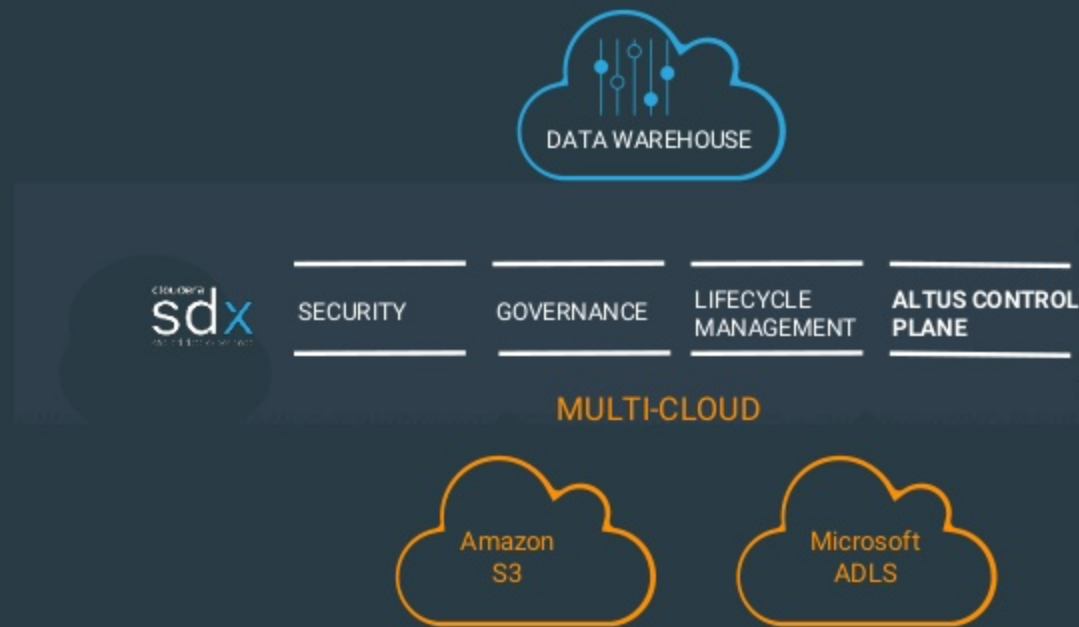


**Stateful Context, Shared Experience**

# CLOUD NATIVE OPTION - ALTUS DW

## MULTI-CLOUD PAAS SOLUTION

- Quick time to value - no software or clusters to manage
- Bring warehouse to the data with zero copy simplicity
- Use your security policies with your data - no proprietary stacks
- Apply enterprise governance to transient workloads
- Shared data experience with SDX
- Optimized for Azure & AWS





# FROM ANALYTICS TO MACHINE LEARNING

Moving from Known Questions on Known Data to Unknown Questions on Unknown Data

## BETTER TOGETHER

Cloudera SDX makes it easy for administrators, BI users, data scientists to work together on a common data set, with consistent data context

DATA  
ENGINEERING



+

DATA  
WAREHOUSE



- Run ETL with Spark or partner tools to ingest and process data at any scale
- Assign permissions and classifications once
- Data, along with all data context, is immediately available in the data warehouse for analytical processing and BI use cases

DATA  
SCIENCE



+

DATA  
WAREHOUSE



- Run data science and machine learning analysis to blend, augment, and score data
- Blended and augmented data, along with all data context, is immediately available to business teams and analysts with unified security and governance

# TOOLS & FRAMEWORKS FOR SUCCESS



# TD BANK: Delivering “Legendary Customer Experience”

## CHALLENGES

Significantly improve customer experience with sentiment analysis, behavioral patterns, and predictive modeling

Current system couldn't handle:

- Centralizing data from thousands of sources
- Demands from increased users and use cases
- Data cost and manageability at scale

## SOLUTION

Modern Data Warehouse for customer marketing, fraud analytics and cybersecurity

- Ingest data from 100+ corporate systems
- Centralized data into “the hands of those that need it much more quickly”
- Significantly reduce storage and management costs

## RESULTS

- 30% reduction in repeat customer complaints
- 90% productivity improvement for analytics projects
- 60% decrease in data management costs
- 98% decrease in per TB storage costs

<https://www.cloudera.com/more/customers/td-bank.html>



# DEUTSCHE TELEKOM: Fraud reduction and customer retention

## CHALLENGES

Improve fraud detection speed to near-real time and respond to network service quality issues before customers notice

Current system couldn't handle:

- Massive volumes of network data - at higher granularity
- Enterprise view of data - machine learning at scale
- Near-real time fraud detection on incoming data

## SOLUTION

Modern Data Warehouse to detect fraud patterns and network problems in real-time before business impact

- Quickly analyze massive streaming data sets
- Enterprise grade reliability and stability with shared data experience (no silos)
- Machine learning and fast analytics - real-time

## RESULTS

- 10-20% reduction in revenue loss by increased fraud detection
- 5-10% decrease in customer churn with increased network quality
- 50% increase in overall operational efficiencies with faster analytics

<https://www.cloudera.com/more/customers/deutsche-telekom.html>



# KOMATSU MINING: Optimize Machine Performance

## CHALLENGES

Create an Industrial IoT (IIoT) solution for optimizing mining equipment utility and build better next-generation products

Current system couldn't handle:

- Scale of IoT data
- Demand for new users and use cases
- 30TB/month data growth

## SOLUTION

Cloud-based IIoT analytics for a full view of mining operations

- Quickly and easily analyze huge volume and variety (time-series, sensor, event, and more) of data
- More use cases and users: "democratizing analytics for different user groups"
- Scale quickly and easily in the cloud

## RESULTS

- 2X Increase in production hours on key equipment
- Design next-generation equipment: environmentally smarter, more productive, at lower cost
- Meet or exceed all KPIs: "Deliver all of the data with less complexity and significant cost savings"

<https://www.cloudera.com/more/news-and-blogs/press-releases/2017-11-15-komatsu-helps-improve-mining-performance.html>

# CLOUDERA DW - PARTING THOUGHTS



Hybrid Optimized



Performance @Scale



Shared Data Experience



Exponential Use Cases, Successful Outcomes

# THANK YOU

<https://www.cloudera.com/products/data-warehouse.html>

**cloudera**

