BENEFITS OF EDW OPTIMIZATION
5 Pain Points
Enterprise Data-warehouse issues that you are facing?

1. Is your EDW becoming **too expensive** to maintain because of **hardware upgrades and increasing data volumes**?
2. Is your EDW becoming a **monolith**, which is **too slow to adapt to business’s analytical requirements**?
3. Can your EDW **scale** linearly to the growing data volumes?
4. Can your EDW handle **unstructured data** & **real-time** requirements?
5. Do you want your EDW to be run in a **Self-Service Mode** for the business users?
You Need Hadoop Optimized EDW
# How Happiest Minds Big Data Offerings Can Help

**Data Lakes**
- Strategy & Roadmap
- Prototyping & Tool Evaluation
- Construction & Go-Live Enablement

**EDW Optimization**
- ELT Offload Architecture
- Datastore, Governance & Security Management
- Self Service BI / Discovery

**Stream Analytics**
- Real-time Ingestion
- Scalable Data Processing & Storage
- Analytics, Dash-boarding & Alerting
Hadoop Optimized EDW

What is it?

• An EDW complemented by Hadoop

• Offload storage / compute from your expensive EDW appliance to Hadoop
  • Identify storage pockets within EDW. Example: Staging database
  • Identify batch processing / real-time processing workflows. Example: ELT/ELT

• Offload Batch Reporting / Self Service BI to Hadoop
  • Identify “Write Once, Publish Many” type of batch reports
  • Identify “Self Discovery” kind of reporting

• Offload Analytics / Machine Learning to Hadoop
How to get started on EDW Optimization?

We recommend to start with a 6 week discovery exercise

IDENTIFY
Identify datamart candidates / reporting use cases to migrate

ARCHITECT
Create Reference Technical Logical & Physical Architecture

EVALUATE
Evaluate / Prototype Hadoop Tools & Technologies

RECOMMEND
Recommend Technology / Governance & Migration Plan
EDW Optimization Phase

MIGRATE
Migrate Candidate EDW Workloads to Hadoop

VALIDATE
Test & Validate results, benchmark performance

REFACTOR
Re-factor to fix issues

RETIRE
Cut-over existing system and move to Hadoop
Benefits of an Optimized EDW

- Reduced Spend on EDW Storage
- Innovate to provide analytics on new age Data Structures
- Improved response times on ELT workload
CASE STUDIES
EDW OPTIMIZATION CASE STUDY

APPAREL RETAILER
Customer wanted to address the following 4 questions?

1. Can Hadoop handle the varied formats of data (CSV, Excel, JSON, XML, Images)?
2. Can Hadoop handle the concurrency of users that we currently support with Teradata?
3. Can Hadoop ingest data in a manner to allow us to meet batch cycle and real-time demands?
4. What Hadoop tools can and should be used to manage data ingestion, data modeling, data at rest and reporting?
Technology Stack – Using MapR Distribution

**Ingest**
- **Source Files** (Avro)
- MapR Streams

**Core**
- **Change Data Capture** (Spark, Hive)
- **Source – Target Schema Converter** (Spark, Hive)
- **Business Transformations** (Spark, Hive)
- **Dimensional Upsert** (Spark, Hive)
- **Dim Lookup for Facts** (Spark, Hive)
- **Reject Processing** (Spark)
  - **Early Arriving Facts** (Parquet)
  - **DQ Rejects** (Parquet)
- **Core E-R Model** (MapR-DB)
- **Data Model (Core & Star)**
- **Type 1 Dimension NOSQL Dataset** (MapR-DB)
- **Atomic Fact Files** (Parquet)

**Publish**
- **Aggregator** (Hive)
- **Flattened Model**
  - **Aggregated Fact Files** (Hive, Parquet)
- **Late Arriving Fact Aggregator** (Hive)
- **Dimension Files** (Hive, Parquet)
- **Downstream Exports** (Hive, Parquet, Text)

**Semantic Views** (Drill)

**Metadata Management**
- **Metadata Export** (Hive)

**File Exporter**
- **Teradata Export** (Sqoop/TDCH)
- **Downstream Export** (Sqoop)
Solution Highlights

1. Use Case Identification
2. Roadmap Consulting
3. Technical Architecture Construction
4. Vendors (HW, Cloudera, MapR) and Tools Evaluation
5. Infrastructure Planning
6. Stewardship & Governance Model Recommendation
7. Training
8. A working prototype for a sample datamart
Case Study: Mobile Analytics for Reliance Jio Media

BUSINESS REQUIREMENT

- Create one stop solution for analytics needs of diverse mobile applications
- Need for a consistent and scalable data-logging framework, reports and analytics for communication services and various digital services in key domains including education, health care, financial services and entertainment.

OUR SOLUTION

- Implement a Big Data solution to handle volume, variety and velocity of the data generated by mobile applications.
- Develop analytic solution to leverage real time, streaming customer data and user experience data.
- Using advanced predictive models such as customer segmentation, decision trees and neural net draw insights to help marketing team devise strategies to retain existing customer and increase customer base.
- Technology Stack: Hortonworks, Kafka, Storm, Spark, Mongodb, Hive

IMPACT

- Reduce customer churn. Improved customer experience
- Increased customer loyalty, satisfaction and revenue
Case Study: Real Analytics on Stream Collection For K12 Education Provider

**BUSINESS REQUIREMENT**

- Scalable solution to support 100,000 messages/sec for 9 millions users.
- Real Time Data Collection, Ingestion and Analytics on Stream data from various sources

**OUR SOLUTION**

- Build data pipeline using Real time messaging system Storm
- Runtime schema resolution and Distributed data store
- Camus Map-Reduce jobs for Batch processing

**IMPACT**

- Get 360 insight by using Batch view of the data
- Collect data from various sources and perform behavioral analytics on student activity
- Feed back analytics results to the business
Experience: Data Lake Architecture for Leading Professional Education & Learning Company

BUSINESS REQUIREMENT

• Multiple Business Units having disparate systems and re-doing same / similar kind of analytics & reporting
• Creation of a Data Lake which pulls in data from, the different Silos and provides a common analytics platform and capabilities

OUR SOLUTION

• Legacy System data was present in databases, which were pulled in
• For new systems consolidated data flowing through Kafka into Azure Blob Storage
• Immediate Data Exploration through – ELK Stack (Elasticsearch / Logstash / Kibana)

IMPACT

• Common Reporting Application minimized the need to re-build same reports for all business units
• Ability to access data through common APIs & direct data mart access provided users to perform in-depth custom analysis
Data Lakes
Data Integration in an Agile fashion is still a big challenge. Unstructured, real-time, high volume data ingestion makes it even more challenging.

SCENARIO #1

Businesses want Self-Service capabilities to perform reporting, data discovery & advanced analytics, rather than spend too much time upfront on design & analysis.

SCENARIO #2

IT Organizations want to scale linearly at affordable computing cost & storage for performing advanced analytics.

SCENARIO #3
DATA LAKES are the way forward
Data Lake
What is a data lake?

• Acts as a **reservoir** for enterprise, social, devices' information.

• **Scalable [ Storage + Compute + Access ]** Data Layer

• **Consumable** by Downstream SQL Users, Analytics Applications, Machine Learning programs, Operational dashboards and BI

• **Governed** by "sufficient" information, departmental policies

• **Secured** by "enterprise-grade" access controls
Evolution of Data Lake
How do you build up to a data lake?

**STORAGE UNITS**

- **Data Storage**
  - Archive Store
    - Application Log Archives
    - Data-warehouse Archives
  - Hadoop/Cloud Exploration
    - Big Data POCs on Hadoop
    - Cloud Feasibility

**COMPUTE CONTAINERS**

- **Single Use Case**
  - Business Problem
    - Recommendation Engine
    - Fraud Analytics
  - System offload
    - Datawarehouse etl offload
    - Operational Analytics

**DATA LAKE**

- **Enterprise Wide**
  - Enterprise, Social, External, Devices, App, IoT data ingestion
  - Self-Service Analytics, Extended BI
  - Interoperability
  - Real-time
  - AI, ML Capabilities
  - Operational Intelligence
6 step approach to building Data Lakes

- **LOAD**
  - Ingest data from enterprise sources
  - Ingest data from external sources
  - Load in data as-is without any schema conversions into Hadoop

- **CURATE**
  - Apply basic schema conversion & DQ Transformations for Self Service consumption
  - Convert to use case based file formats

- **GOVERN**
  - Manage resource, access, security & metadata
  - Manage data retention & hot/cold data strategies
  - Manage Quality & Privacy Policies

- **FAMILIARIZE**
  - Mobilize the organization on Data Lakes adopting a few business use cases
  - Create demos to showcase the benefits of Data Lakes, before embarking a full blown project

- **USE**
  - Once the utility is proved, start leveraging 7 using the data lakes for all the benefits highlighted

- **REFACTOR**
  - Refactor the Data Lakes, based on new use cases & technologies
  - Add new use cases
  - Add the right tool for the job
Logical Architecture

- Robust Metadata discovery, Governance & Security Policies
- Easy to use Self Service Capabilities
- Linear Scalable Storage, Compute & Access Layers
- Cost-efficient Infrastructure
- Fit for purpose tooling rather than one-size-fits-all approach
Our Data Lake Consulting Methodology
What is our methodology to recommend data lake initiatives?

TOP-DOWN

Define Use Cases
Understand Information Policies
Understand Governance Requirements

RECOMMENDATIONS
- Prioritized Use Cases
- End State Architecture
  - On-premise vs Cloud
  - Open source vs Vendor
- Tool Recommendation
  - Ingestion tools/frameworks
  - Distribution recommendation
  - SQL tools
  - Metadata tools
- Governance & Security Recomm.
- Migration Roadmap
  - Migration Strategy
  - Source Adoption plan
  - Project plan
  - Effort & Cost Estimates
  - Infrastructure Estimates
- Proof-Of-Concept (Optional)

BOTTOM-UP

Understand Candidate Data Sources
Assess their existing Data Architecture
Perform Gap Analysis & Assess DL Migration Readiness
Partnerships & Alliances
STREAMING ANALYTICS
1. Are you constrained by data workflows that handicaps your business from taking faster decisions?

2. Do you run a business where the value of data decreases exponentially as it ages? (Last 10 minutes of data is more valuable than Last 2 weeks of data)

3. Have you missed revenue opportunities or incurred losses because your systems didn’t proactively alert at the right time?

4. Do you want your decision support systems to identify outliers in less than 60 seconds of occurring?
You Need
Real Time Streaming Analytics Platform
Platform Logical Architecture
A reference architecture for a streaming data platform

Sources
Real-time Sources
- Avro Serialized
  - KAFKA
    - RAW DATA SERVICES
    - TRANSFORM DATA SERVICES
    - EXPORT DATA SERVICES
- Flume
  - AGENT 1
  - AGENT N

Log Data
- REST

SPARK STREAMING
- STORE
  - AGGREGATOR
  - ARCHIVE
  - EXPORT
  - TRANSFORMER

HADOOP
- HDFS
  - Time Series NoSQL Datastore (HBase)
  - Real time Aggregated Store (HBase)
  - OLAP Store (Hive)
  - Master Data Metadata (HBase)

REAL TIME DASHBOARDS
- FEEDBACK SERVICES

OPERATIONAL REPORTS
- RDBMS (MySQL)
  - ANALYTICAL TOOLS (R/PySpark)
  - ADMIN SERVICES
CASE STUDIES
Case Study: Mobile Analytics for Reliance Jio Media

**BUSINESS REQUIREMENT**

- Create one stop solution for analytics needs of diverse mobile applications
- Need for a consistent and scalable data-logging framework, reports and analytics for communication services and various digital services in key domains including education, health care, financial services and entertainment.

**OUR SOLUTION**

- Implement a Big Data solution to handle volume, variety and velocity of the data generated by mobile applications.
- Develop analytic solution to leverage real time, streaming customer data and user experience data.
- Using advanced predictive models such as customer segmentation, decision trees and neural net draw insights to help marketing team devise strategies for retain existing customer and increase customer base.
- Technology Stack: Hortonworks, Kafka, Storm, Spark, Mongodb, Hive

**IMPACT**

- Reduce customer churn. Improved customer experience
- Increased customer loyalty, satisfaction and revenue
Case Study: Real Analytics on Stream Collection For K12 Education Provider

BUSINESS REQUIREMENT

• Scalable solution to support 100,000 messages/sec for 9 million users
• Real Time Data Collection, Ingestion and Analytics on Stream data from various sources

OUR SOLUTION

• Build data pipeline using Real time messaging system Storm
• Runtime schema resolution and Distributed data store
• Camus Map-Reduce jobs for Batch processing

IMPACT

• Get 360 insight by using Batch view of the data
• Collect data from various sources and perform behavioral analytics on student activity
• Feed back analytics results to the business
Case Study: Streaming Analytics solution for a large media & entertainment company

**BUSINESS REQUIREMENT**

- Scalable solution to support more 100,000 messages/sec
- Real Time Data Collection, Ingestion and Analytics on Stream data from various sources

**OUR SOLUTION**

- Build data pipeline using Real time messaging system - Kafka, Spark Streaming, Timeseries Data Store (OpenTSDB) & Grafana
- Display real-time operational monitoring dashboard

**IMPACT**

- Get immediate insights into outliers on payment declines
- Provide actionable insights on segments causing revenue loss and trends around it
Appendix
About Happiest Minds

Next Generation Digital Transformation, Infrastructure, Security and Product Engineering Services Company

- Launched in August 2011
- Raised Series A Funding of $63Mn USD
- Our Investors: J.P. Morgan, Intel Capital, Ashok Soota
- 2400+ People
- 170+ Customers
- 16 Cities
- 8 Countries

Logos:
- Deloitte Technology Fast50
- 2017 India Digital Transformation Awards
- IAOP The Global Outsourcing 100
- India's Best Companies To Work For 2016
- NASSCOM Tech Series 2017 Big Data & Customer Analytics
- ThreatVigil
Our Business

- Mobility
- DevOps & RPA
- Software Defined Networking / NFV
- Big Data & Adv. Analytics

- IoT
- Cloud
- BPM & Integration
- Security

- BFSI
- Retail
- CPG
- HiTech
- Mfg/Industrial
- Travel & Hosp.
Digital Transformation & Enterprise Solutions

Business Efficiency
- Business Agility
- Cost Optimization
- Informed Decisions
- Increased Productivity
- Innovation

Digital Transformation

Customer Experience
- Personalized Interactions
- Dynamic & Aware Apps
- Deeper Engagement
- Channel Flexibility
- Rich User Experience
Analytics

Big Data Engineering
- MPP to NoSQL migration
- Data Lakes
- Real time streaming frameworks

DW on Cloud
- HDFS Migration
- Advanced Viz.
- BI/DW

Data Science
- Problem identification
- Machine learning models
- RT model implementation

AI/ Cognitive
- Deep Learning
- Chat Bots
- GPU Computing

Solutions

Partnerships

Retail
- 200+ SMEs
- 25+ Customers
- 10+ Partnerships

Banking

Healthcare

Manufacturing

Telecom

Real Estate