<table>
<thead>
<tr>
<th>CHALLENGES</th>
<th>SOLUTION</th>
<th>BENEFITS</th>
<th>QUICK FACTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>DATA WAREHOUSE OPTIMIZATION WITH MAPR</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
BUILDING A MODERN DATA ARCHITECTURE: OPTIMIZING FOR COST AND PERFORMANCE

Traditional enterprise architecture promotes the division between two distinct workloads—applications and analytics. Applications help run the business and analytics help understand it. These two workloads always had their separate technology stacks. The analytics side of the architecture was (and to some extent still is) dominated by data warehouses. Increasing data volumes, new types of data formats, and emerging analytics technologies such as machine learning have forced a rethink of data warehouse centric data architecture and presented following challenges.

Ever increasing cost
Expensive licensing models combined with the massive volumes of data being created means the cost of data warehousing is significantly rising.

Inability to scale-out
Data warehouses cannot scale-out linearly using commodity hardware. Buying new expensive hardware is straining IT budgets.

Unused data driving cost up
70% of data in DW is unused, i.e. never queried in past 1 year.

Misuse of CPU capacity
Almost 60% of CPU capacity is used for ETL/ELT. 15% of CPU consumed by ETL to load unused data. This affects performance of queries.

Inability to support non-relational data
Designed for relational data, data warehouses are not suitable for unstructured data coming from sensors, logs, devices, social media etc.

Inability to support modern analytics
DWs do not support modern analytics technologies such as machine learning and stream processing.
OPTIMIZING YOUR DATA WAREHOUSE: 3 STEP SOLUTION FROM MAPR

Select the ideal offload candidates
MapR experts will help you select the data and ETL workload ideal for offload. Keep the frequently queried data in the data warehouse. Select unused data (often up to 70% of the data in the DW) for offloading into the MapR Converged Data Platform, which will serve as the data lake. Many of the CPU intensive ETL workloads can be offloaded to the MapR data platform, especially the ones related to unused data. Apache Spark can take over these ETL workloads on the MapR Platform.

Build the data pipeline
Data migration can be performed using batch methods using NFS or Sqoop or real-time methods using tools such as Kafka Connect for MapR-ES. Many data warehouses provide connectors for Hadoop that help to simplify the migration. Upon migration, the data can be stored in MapR-DB or Hive tables, or Parquet or Avro files depending on requirements.

Deliver the data to the stakeholders
Utilize SQL engines such as Apache Drill, Hive, or Spark SQL to deliver data to traditional BI tools. You can continue using your favorite BI tools such as Tableau, Qlik, Microstrategy etc. The existing BI teams can continue querying the offloaded data using SQL. This solution ensures smooth and continuous operation for BI teams.
DATA WAREHOUSE OPTIMIZATION REFERENCE ARCHITECTURE

**ETL PROCESSING**
MapReduce, Spark, Pig

**SQL QUERY ENGINE**
Drill Interactive SQL
Hive SQL Batch Processing

**MAPR-ES**
Real-time Events

**MAPR-XD**
Files (Parquet, Avro, CSV, Text)
Logs, Blogs, Images

**MAPR-DB**
Structured Data, Time-series Data, JSON Documents

**OLTP, ERP, CRM**
**DOCUMENTS**
**CSV**
**TEXT FILES**
**SOCIAL MEDIA**
**CLICKSTREAM**

**NFS DIRECT ACCESS**
**KAFKA API REAL-TIME EVENTS**
**SQOOP BATCH INGEST**
**FLUME INGEST LOG DATA**

**OFFLOAD / RELOAD**
**NFS DIRECT ACCESS**
**KAFKA API REAL-TIME EVENTS**
**SQOOP BATCH INGEST**

**DATA MARTS**
**DATA WAREHOUSE**

**BUSINESS INTELLIGENCE**
**ANALYTICS**
**REPORTING**
MAPR CONVERGED DATA PLATFORM: KEY FEATURES

The MapR Converged Data Platform integrates analytics powered by Apache Drill, Apache Spark and Apache Hadoop with real-time database capabilities, global event streaming, and scalable enterprise storage to power a new generation of big data applications. The MapR Platform delivers enterprise-grade security, reliability, and real-time performance, while dramatically lowering both hardware and operational cost of your most important applications and data.

Interactive SQL analysis
Apache Drill on MapR platform allows you to use ANSI SQL to query any data. BI teams can continue using SQL and the same BI tools.

Multi-temperature storage
MapR provides multi-temperature storage. Store your hot, warm, and cold data within MapR on hardware of your choice, further optimizing for cost.

Streaming for real-time insights
MapR-ES allows you to bring data for analysis as soon as the data is created. In contrast, legacy DW solutions are batch oriented.

Multi-tenant big data platform
MapR is the only big data platform that provides multi-tenancy on the data placement level, helping you meet the regulatory requirements.

Enterprise grade data governance
The MapR platform provides enterprise grade security, auditing, and lineage to meet your data governance and regulatory compliance needs.

Converge SQL and Machine Learning
Single platform for storage, database, and streaming. Your choice of compute engine on top (Spark, Hadoop, SQL, Machine learning).
### Mapr Converged Data Platform: Key Business Benefits

Data Warehousing optimization projects are typically driven by cost reduction initiatives. However, you also want to ensure business continuity, by making sure the transformation is transparent for most of the users of the data. This is also a great opportunity to build a data platform for the future that will bring all data into play, and allow you to leverage up and coming data analysis technologies such as machine learning.

<table>
<thead>
<tr>
<th><strong>Reduce TCO of data analysis</strong></th>
<th><strong>Leverage existing skillsets</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sharply reduce the cost of data management and analytics. The cost saving can be utilized towards revenue generating innovations.</td>
<td>BI teams can continue using familiar tools such as Tableau, Qlik, Microstrategy etc. on both Data Warehouse and the MapR data lake.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Maximize value of current investment</strong></th>
<th><strong>Leverage hybrid deployment model</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Increase available “headroom” and avoid or minimize new CapEx. Improve performance of your existing data warehousing assets.</td>
<td>MapR provides single global namespace to help you create a homogeneous data fabric across on-premises and cloud hosted data.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Answer new questions</strong></th>
<th><strong>Get results faster</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Use analytics tools unavailable in legacy data warehousing (Drill, Parquet, Spark, Machine Learning, others).</td>
<td>The MapR Quick Start Solution for Data Warehousing will help you get value for optimization project within 5 weeks.</td>
</tr>
</tbody>
</table>

---

**$10M**

IN SAVINGS PER 1 PB OVER 5 YEARS

**53%**

LOWER CAPEX FOR THE COMBINED DW + MAPR SOLUTION

**90%**

LOWER CAPEX FOR MAPR VS LEGACY DW
BEYOND COST REDUCTION: MAPR CAN HELP YOU TRANSFORM YOUR BUSINESS

IDC conducted interviews with organizations using MapR as a Big Data platform to understand how they are leveraging it to drive their businesses with Big Data. Besides $4X$ ROI on MapR investment, key business benefits from the IDC survey findings include $19.44$ million average business benefits over three years, $31\%$ higher data scientist productivity, $39\%$ higher application developer productivity, and $42\%$ lower cost of operations than alternative Big Data solution. MapR has a proven track record of helping customers build the next generation data platform for business transformation.
**SUMMARY**
Increasing data volumes, new types of data formats, and emerging analytics technologies such as machine learning have forced a rethink of data warehouse centric data architecture.

**CHALLENGES**
- Ever increasing cost for existing Data Warehousing
- Inability to scale-out
- Unused data driving cost up
- Misuse of CPU capacity for ETL
- Inability to support non-relational data
- Inability to support modern analytics

**SOLUTION**
3 Step solution from MapR:
- Select the ideal offload candidates
- Build the data pipeline
- Deliver the data to the stakeholders

**KEY FEATURES**
- Interactive SQL analysis
- Multi-temperature storage
- Streaming for real-time insights
- Multi-tenant big data platform
- Enterprise grade data governance
- Converge SQL and machine learning

**BENEFITS**
- Reduce TCO of data analysis
- Maximize value of current investment
- Answer new questions
- Leverage existing skillsets
- Leverage hybrid deployment model
- Get results faster with the MapR Quick Start Solution

**QUICK FACTS**
- 3 Step Process
- 5 Weeks Time-to-Value
- $10M in savings per 1 PB over 5 years

---

**LEARN MORE** Data Warehouse Optimization with MapR
**CONTACT US** info@mapr.com
**TRY MAPR** Download