Identifying and Exploiting the Keys to Digital Transformation

Jack Norris
SVP Data & Applications
MapR Technologies
Putting Things into Perspective

INVESTMENT IN NEXT-GEN VS. LEGACY TECHNOLOGIES FOR DATA

$ (millions)


Next-Gen technology investment Legacy technology investment Total $ growth of IT market

Source: IDC, Gartner; Analysis & Estimates: MapR
Next-gen consists of cloud, big data, software and hardware related expenses

90% of data is on next-gen technology by 2020
A Once-in-30-year Re-platforming of The Enterprise

<table>
<thead>
<tr>
<th>New Applications</th>
<th>Existing Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open Source</td>
<td>Analytic Innovations</td>
</tr>
<tr>
<td>On Premise</td>
<td>Private Cloud</td>
</tr>
<tr>
<td>Heterogeneous Hardware</td>
<td></td>
</tr>
</tbody>
</table>

- Critical infrastructure for next-gen applications
- Data platform enabler for required Speed, Scale, & Reliability
Data Warehouse Offload: Cost per Terabyte Comparisons

- Augmenting legacy infrastructure can save over $1.5K or 79% in annual op ex per TB and slash the cap ex per TB by $4.8K or 95%

- Legacy: Typical net Exadata pricing with rack, storage server, and DB bundle software
  - Assumes 100% use of available storage
Platform for Innovation
Largest Biometric Database in the World
The Keys to Digital Transformation

1. Data Convergence
Legacy Business Platforms Are Two Islands

- IT must integrate all the products
- Inability to operationalize the insight rapidly
- Can’t deal with high speed data ingestion and processing
- Scale up architecture leads to high cost
Traditional Applications Dictate Data

Application

Data

ETL process

Specialized schema

Security
Applications Dictate Data – Results in Silos
Before

- Complex and Slow
- Multiple Versions of the Truth
- Difficult Governance
- High TCO

Converged

- Real-Time Data to Action
- Single Data Copy
- Easy Governance
- Low TCO – Scales Horizontally

HTAP (Hybrid Transaction/Analytical Processing) – Gartner Group 2015
Point Products Create Crisis of Complexity

EXPENSIVE TO STITCH | FRAGILE | LIMITATIONS FOR SPEED, SCALE, RELIABILITY
Yield Management Optimization
Big Data is Generated One Event at a Time

“time”: “6:01.103”,
“event”: “RETWEET”,
“location”:
“lat”: 40.712784,
“lon”: -74.005941

“time”: “5:04.120”,
“severity”: “CRITICAL”,
“msg”: “Service down”

“card_num”: 1234,
“merchant”: "Apple",
“amount”: 50
The Keys to Digital Transformation

2. Stream Processing
Event-based Data Flows

- web events
- machine sensors
- Biometrics
- Mobile events
- etc…
Streams Enable Real-Time Applications

- Failure Alerts
- Trending now
- Real-time application & network monitoring
- Web Personalized Offers
- Supply Chain Optimization
- Ad optimization
- Real-time Fraud Detection
- Ad optimization
- Supply Chain Optimization
- Web Personalized Offers
- Trending now
- Real-time application & network monitoring
- Failure Alerts
What’s a Stream?

A **stream** is an unbounded sequence of events carried from a set of producers to a set of consumers.

Producers and consumers don’t have to be aware of each other, instead they participate in shared **topics**. This is called publish/subscribe.
Events are delivered in the order they are received, like a queue.
Unlike with a queue, events are persisted even after they’re delivered.
Flexible Communication Modes

One to One

Many to Many

Orders

Clicks

Consumer Group
Transforming the Health Care Ecosystem

“The Stream is the System of Record”
– Brad Anderson
VP Big Data Informatics

JSON DB (MapR-DB)
Graph DB (Titan on MapR-DB)
Search Engine (Elastic-Search)

Electronic Medical Records

“Electronic Medical Records”
– Brad Anderson
The Keys to Digital Transformation

3. The Importance of Agility
Three Dimensions To Convergence

Data Agility
• Unified Files, Tables, Streams
• Support for schemas that change
• Multi-modal support in a DBMS

Application Agility
• Microservices-based development
• Streaming applications in real time
• Machine learning

WEB-SCALE STORAGE
DATABASE
EVENT STREAMING

High Availability  Real Time  Unified Security  Multi-tenancy  Disaster Recovery  Global Namespace

Infrastructure Agility
• Multi-dimensional elasticity
• Global multi-cloud
• Container apps with data persistence
Distributed Processing Across Locations

Distributed processing across clusters, data centers and public and private cloud environments

Flexible processing where change is the norm

Supports global apps that can scale arbitrarily
Cloud Processing Does Not Solve Crisis Of Complexity
Core Architecture Requirements for Distributed Processing

- Location awareness
- Continuous, Coordinated, Data flows
- Strong consistency by default
- Global single namespace
- Omni-directional replication
Data Replication

- Dynamically move data between sites
- Streams, replication
- Use Cloud resources as a disaster recovery site
- Mirroring
Application Execution

- Intelligently place workloads in the geography where data is local
- Balance needs of high-throughput and low-latency applications
- Support distributed database execution
Distributed processing (device, edge, cloud, on premise)

Location driven by
- Data gravity
- Safe harbor
- Costs
- Resource availability
Extending Convergence to the Edge

• **Small footprint**
  – 3-5 Node Cluster (limited disk capacity)
  – We will consider 1 node in future

• **Expand into edge locations**
  – E.g., oil rigs, hospitals, IoT aggregators, remote offices

• **Benefits**
  – Reliably consolidate data from edge sites to core data center
  – Gain insight at edge using analytics
  – Simplify management using familiar tools
  – Increased data protection
    • Replicate edge data to core with secure communications
    • Secure all data in MapR with strong security controls
Oil Exploration and Production Example
Oil Exploration and Production Example

1000s of Oil & Drill Sources

Manual Process:
- Source 1
- Source 2
- Source 1000
- Time to insight (48 hrs)

Before Edge

Automated Process:
- Source 1
- Source 2
- Source 1000
- Will do Pre Processing locally + at Core (Custom App + Down Sampling)
- Time to insight (<2 hrs)

After Edge

Houston MAPR Core Cluster
Connected Car Example

By 2020, more than 250 million vehicles will be connected globally
- Gartner
Car Company Test & Dev Case Study

Before Edge

1000s Test Cars running in various Conditions 24/7
1-5TB/Day

Manual, Scripts, Req High Bandwidth Network

All Analysis only done at Core

Can do Pre Processing locally + at Core (Custom App + Test Feedback Loop)

Automated Process

After Edge

Time to insight (24 hrs)

Car Test Analysis App

Time to insight (< min)
Medical Device Company Example
Medical Device Company Example

100s of Medical machines in each hospital

Source 1

Source 2

Source 100

Manual, Scripts, File Sync SW Process

Time to insight (12 hrs)

Before Edge

100s of Hospitals

Machine Learning Diagnosis App

Automated Process

Will do Pre Processing locally + at Core (Custom App + Anonymizations)

Time to insight (<15 mins)

After Edge

100s of Hospitals

Machine Learning Diagnosis App
Clouds And Containers

80% say Docker is part of cloud strategy

60% plan to use Docker to migrate workloads to cloud

41% want application portability across environments

35+% want to avoid cloud vendor lock-in

Source: 2016 Docker Survey
Challenge With Containers

• Organizations using containers to improve resource utilization, increase developer efficiency, deploy microservices

• Most of containerized applications today are stateless
  – If destroyed and re-deployed, no data is lost and no services are disruptive (ex: simple web front-ends)
  – Sharing state across many nodes is difficult and/or cost-prohibitive

• Container environments don’t provide the reliable storage needed to support stateful applications
Agility for Docker

Flexible
Hybrid/Multi Cloud Infrastructure
Agility for Docker

Flexible Hybrid/Multi Cloud Infrastructure
Agility for Docker

Flexible Hybrid/Multi Cloud Infrastructure

Web-Scale Storage  
Database  
Event Streaming  

MapR-FS  
MapR-DB  
MapR Streams  

High Availability  Real Time  Unified Security  Multi-tenancy  Disaster Recovery  Global Namespace
Data Services For Containers To Support Stateful Apps

- Pre-built, certified container image for access data from converged platform
- **Secure** authentication at container level, secure connection
- **High performance**
- **Extensible** support for application layers
- Available in Docker Hub, Dockerfile for **customizability**
What Does this Mean For Development?

APPLICATION & DATA MODELS RADICALLY CHANGING

Each application solved one problem and created its own data type

Diverse data assets must be accessible from anywhere by microservices
Microservices Architecture

Processing Services

Processing Services

Message Broker

Message Broker
Microservices Architecture

Processing Platform

Streaming Data Platform

Message Broker
Converged Microservices Architecture

- Simplifies microservices development
- Developers have direct access to files, databases, and documents integrated with the publish-and-subscribe framework
Data Protection and Fast Recovery

• Snapshot captures all data, including data-in-motion and data-at-rest.

• Consistent point-in-time recovery for the entire process

• Simplifies recovery, administration and development
Appeal For Developers

Flexibility
Agility
Simplicity
Real Time
Appeal For Architects And Administrators

Unifies security, data protection, disaster recovery
Simplifies administration
Avoids cluster sprawl
Existing Environments – Complex Silos
Road to Digital Transformation