Cloud computing and big data

September 27, 2016

Ben Sharma | CEO
ben@zaloni.com
• Award-winning provider of enterprise data lake management solutions:

**BEDROCK** Integrated data lake management platform

**MICA** Self-service data preparation

• **Data Lake Design and Implementation Services:**
  - POC, Pilot, Production, Operations, Training

• **Data Science Professional Services**
Why cloud now?

From IDC Research:

“By 2018, at least half of IT spending will be cloud-based, reaching 60% of all IT infrastructure”

“By 2018, cloud becomes a preferred delivery mechanism for analytics, increasing public information consumption by 150%”
## Why are companies moving to a cloud-based platform

<table>
<thead>
<tr>
<th>Infrastructure Drivers</th>
<th>Data Locality</th>
<th>New Requirements</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Infrastructure agility</td>
<td>• Data Gravity</td>
<td>• Lot of data is generated externally</td>
</tr>
<tr>
<td>• Cost</td>
<td>• Compliance and regulatory requirements (international)</td>
<td>• Need to handle all types of data – Structured, unstructured, images, etc.</td>
</tr>
<tr>
<td>• Compute and storage elasticity</td>
<td>• Keep data close to where it is generated</td>
<td>• Latency and Currency</td>
</tr>
<tr>
<td>• Heterogeneous compute and storage platforms</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Converged architectures for various workloads</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
IoT volume driving to cloud adoption

By 2020 there will be **5.4 BILLION**

Connected devices¹, like smart meters and connected cars — This is the Internet of Things. And it’s going to be big…

Source: ABI Research

Cloud computing required to provide the virtual infrastructure needed to process enormous volume of data from the IoT

---

¹ Source: ABI Research
Hadoop deployment trends

- On-Premises: 32%
- Cloud Plus On-Premises: 29%
- Cloud Only: 23%

Gartner’s Sept 2015 report: Hadoop Expansion Boosts Cloud and Unsupported On-Premises Deployment
Cloud big data use case: Real-time data processing

Fleet

Data Collection

Idle Time Calculation

Idle Time reporting

On-board Unit

Data Collectors

Ingestion

Streaming Analytics

Dispatchers

Data-driven Apps

Spark Streaming

Apache HBase

AngularJS
Cloud Data Lake options

• Storage – Block, object and file level abstractions, with different degrees of redundancy, availability and consistency guarantees, and cost considerations.

• Compute - A variety of compute server types are possible, optimized for different types of memory and processing requirements depending on the workload.

• Cloud native services – Higher levels of platform abstractions such as cloud provider managed Hadoop clusters, managed databases, warehouses, messaging services, etc.

• Data Management, Governance, Entitlements and Security
Cloud Data Lake Maturity model

Lift and Shift
- Replicate on-premise Data Lake in the cloud

Cloud Native features
- Leverage Object stores, Transient compute platforms, Messaging systems

Multi and Hybrid Cloud
- Abstraction over multiple clouds, consistent Data Management and Governance
Patterns and Anti-patterns

- **Patterns:**
  - Implement Data Lake in the cloud using elastic compute and cloud optimized storage
  - Use Data Lake provided as a cloud service that is managed and optimized by the cloud provider
  - Data pipelines with processing components decoupled by queuing services
  - Leaving the heavy lifting to cloud provider services, example, for elastic clusters, streaming, analytics and machine learning
  - Using cloud storage rather than ephemeral storage with data lifecycle management
  - Real time processing with event driven architectures for streaming data
Patterns and Anti-patterns

- Anti-Patterns:
  - Fork lift migration of on-premise Data Lake to the cloud.
  - Unmanaged, unmonitored, long term usage of resources such as persistent on-demand compute instances.
  - Dedicating cloud resources for service peaks rather than using auto scaling cloud services
Governance considerations within cloud/hybrid environments

- Repeatable Ingestion of vast amounts of data from a wide variety of sources and formats (streaming, files, custom)
- Data visibility across hybrid cloud environments with proper security and access control. Data Masking, and Encryption of sensitive data
- Need to capture operational metadata implicitly during ingestion and processing. Metadata persistent across cluster instances
- Reusable Managed Data Pipelines for Processing: Validation, Standardization, Enrichments
Assessing Cloud Data providers

- Data Lake on IaaS with bare metal or virtualized infrastructures.
- PaaS layers - managed data platforms that include various options for event based data ingestion, data processing and serving layers.
- Several cloud providers are also starting to offer Analytics as a Service with Machine Learning offerings built on top of their IaaS and PaaS layers.
- Geographical coverage due to any local in-country data requirements.
- Cost, TCO for Cloud Data Lake
Cloud options in the context of big data and data science

**Cloud Providers**

- Analytics: Machine Learning, Cortana, Cloud Machine Learning
- Streams: AWS Lambda, Streaming Analytics, Dataflow
- Platform: Amazon EMR, Microsoft, Dataproc
- IaaS: Amazon Web Services, Microsoft Azure, Google Cloud Platform

**Hadoop Ecosystem**

- OR: Spark, H2O.ai, R
- OR: MLlib
- OR: Spark, STORM, Flink
- OR: cloudera, Hortonworks, MAPR
The Next Generation

DATA LAKE

Data Lake 360°

ZALONI
THE DATA LAKE COMPANY

MICA
SELF-SERVICE DATA PREPARATION

Bedrock
DATA LAKE MANAGEMENT AND GOVERNANCE PLATFORM
Visit Booth #644 for these giveaways!

FREE T-SHIRT!

Demo and FREE copy of book "Architecting Data Lakes"

Speaking Sessions:

**Cloud Computing and Big Data**
Ben Sharma, CEO and Founder, Zaloni
Tuesday, 9:30 a.m. – 1B 01/02

**Building a Modern Data Architecture**
Ben Sharma, CEO and Founder, Zaloni
Wednesday, 2:05 p.m. – 1E 09