Azure Data Lake
How to organize

Jan Cordtz, Microsoft Denmark
jcordtz@Microsoft.com
Cloud Solution Architect
More certifications than any cloud provider

**GLOBAL**

- ISO 27001
- ISO 27018
- ISO 27017
- ISO 22301
- CSA STAR Certification
- CSA STAR Attestation

**INDUSTRY**

- PCI DSS Level 1
- CDSA
- MPAA
- FACT UK
- Shared Assessments
- FISC Japan
- HIPAA / HITECH Act
- HITRUST
- GxP 21 CFR Part 11
- MARS-E
- IG Toolkit UK
- FERPA
- GLBA
- FFIEC

**REGIONAL**

- Argentina PDPA
- EU Model Clauses
- UK G-Cloud
- China DJCP
- China TRUCS
- Singapore MTCS
- Australia IRAP/CCSL
- New Zealand GCID
- Japan My Number Act
- ENISA IAF
- Japan CS Mark Gold
- Spain ENS
- Spain DPA
- India MeITY
- Canada Privacy Laws
- Privacy Shield
- Germany IT Grundschutz workbook

**Trusted**
Always been there but growing “rapidly”

Been there for a long time (BI) but getting much more advanced – Machine Learning/AI

“New” kid on the block
- Unlimited compute/storage
- Fast deployment
- Pay-as-you go
- Many services
## Applications

| WordPress | Drupal | OpenShift Enterprise | Cloud Foundry |

## DevOps

| Chef | Ansible | Puppet | Jenkins |

## Frameworks

| .NET | Node.js | Python | Java | PHP | R |

## Databases & middleware

| Cloudera | Hortonworks | MariaDB | Cassandra |

## Containers

| Docker | DC/OS | CoreOS | Kubernetes |

## Infrastructure

| DC/OS | Red Hat | ConsenSys | Blockstack | Ubuntu | Linux | Docker | FreeBSD |
The Azure Data Lake
The technics and organization
Business needs

- Governance
- Organize
- Common understanding of data
- Trial: Error/Proceed
- Hot/Cold path
- No specific technology
- Flexible economy
- Hybrid

Mode 1
- Datawarehouse
- Reporting

Central platform
“Data Lake” / “Data Bank”

Mode 2
- IOT
- Machine Learning
- Analytics

Selfservice
- Dashboard
- Business Intelligence

Flexible economy - Hybrid

Trials: Error / Proceed - Hot/Cold path - No specific technology - Common understanding of data - Organize - Governance
Built on Open Standards

Built on YARN

Store lets all HDFS compliant analytic applications connect to it like Hortonworks, Cloudera, and MapR.

Microsoft HDInsight is 100% Apache Hadoop.

Microsoft continues to contribute tens of thousands of code and engineering hours to open source.

Built using open standards
A databank

- Operational System A
- Operational System B
- Operational System X

Data Lake

- Data Ingestion
- Archive storage
- External Data

Fixed Sizeable - IAAS Economics
- Dynamic Sizeable - PAAS Economics

- SQL
- DW
- Cosmos DB
- Cube
- DataMart
- Machine Learning
- Data Bricks
- Whatever Apps
- Whatever Apps
Data storage

- Azure
  - Machine Learning services
  - Data Science Virtual Machine
  - Machine Learning Studio
  - DataBricks
- Blob
- Data Lake
  - Cloudera*
  - HDInsight
  - HortonWorks*
  - MapR*
- DB
- Rest
- Files
- SQL DB
- SQL DW
- Microsoft SQL (General)
- PostgreSQL (GIS)
- MySQL (LAMP/PHP)
- Key-Value (Table)
- Column-Value (Cassandra)
- Documents (MongoDB)
- Graph (Tinkerpop)
- NoSQL
- SQL
Storage – from a functionality point of view

File storage
- Windows like
- SMB protocol
- No Functionality
- Some limits

Data Lake
- Linux like
- HDFS protocol
- Some functionality
- No limits

Database
- Relational DB like
- SQL protocol
- Rich functionality
- DB Limit

Cube
- Highly specialized functionality

Functionality and cost
- ETL Data Factory
- Data Factory
Principal regarding the Organization

- Is very simple to use for an end-user/application (=flat file/csv file)
- Is as cost-effective as sensible/possible.
- Do not compromise security.
- Fits well into a DevOps scenario
- "Automatic" meta-tagging
- Have a well-defined path for the information needed to be able to support an effective auditing and logging process.
Organizing the Azure Data Lake

**Azure Data Lake**

- **Landing Zone**
  - Landing Zone System Account(s) - read/write
  - Work System Account(s) - read/write

- **Work**
  - Work System Account(s) - read/write

- **Transform**
  - Transform & Anonymize

- **Publish A**
  - Folder per “area”
  - Users in Groups
  - Read/Write

- **Publish B**
  - Folder per “area”
  - Users in Groups
  - Read/Write

- **Publish X**
  - Folder per “area”
  - Users in Groups
  - Read/Write

- **Analytics**
  - “All data”
  - Users in Groups
  - Read/Write

- **Data Catalog**

- **Data Ingestion**
  - Copy

- **Archive**

- **Copy**
Data Ingestion

Push/Pull

Hot/Cold Path

Standardization

Items like: Date formats (yyyyMMdd), number formats (, or .)

Validation

"Is the content you are coming with in accordance with what we have agreed"

"Gatekeeper"

"Are you allowed to enter?"

Examples

SSIS, Event Hub, Data Factory......

Database, FTP, File Storage......

Firewall, AD control......
Thank you