IBM Cloud Data Services - An Overview

Avijit Chatterjee, Ph.D.
STSM, Member of IBM Academy of Technology
achatter@us.ibm.com
@ChatterAvijit
Agenda

- Cloud Solution Considerations
- Cloud Data Services
  - Cloudant
  - dashDB
  - DB2 on Cloud
  - Compose
  - BigInsights on Cloud
  - Spark
  - DataWorks Forge
  - Watson Analytics
- CDS Integration Examples
Agenda

- **Cloud Solution Considerations**
- **Cloud Data Services**
  - Cloudant
  - dashDB
  - DB2 on Cloud
  - Compose
  - BigInsights on Cloud
  - Spark
  - DataWorks Forge
  - Watson Analytics
- **CDS Integration Examples**
Through 2020, the most common use of cloud services will be a hybrid model combining on-premises and external cloud services.”

Gartner, Cloud Computing Innovation Key Initiative Overview, July 2014
Clients Are Aggressively Moving Workloads to Cloud

- DevOps
- Disaster Recovery
- Collaboration
- Web Applications
- e-Commerce
- Customer Service
- Mobile
- Front Office / Desktop
- Social Business
- 3rd Party Applications
- ERP / CRM
- High Performance Computing
- Compute Workloads
- Business Processes (e.g. Expense Reporting)
- Storage Workloads

- Archive
- Database Workloads
- Risk & Compliance
- Big Data & Analytics
- Information Intensive Applications
- Batch Processing

- Mature Workloads
- Isolated Workloads
- Applications with Sensitive Data
- Applications with Complex Processes & Transactions
- Highly Customized Applications
- Not yet virtualized applications

- Not Ready for Cloud
- May be ready for Cloud
- Moving to Cloud
New and Existing Workloads Use Cloud Differently

Cloud Enabled

- Scalable
- Virtualized
- Automated Lifecycle
- Heterogeneous Infrastructure

Existing Middleware Workloads

Compatibility with existing systems

Cloud Centric

- Elastic
- Multi-tenant
- Integrated Lifecycle
- Standardized Infrastructure

Emerging Platform Workloads

Exploitation of new environments

Compatibility with existing systems

Exploitation of new environments
New Markets and New Buyers

One of the biggest opportunities in the cloud is reaching new clients and new buyers within existing clients

Dealing with these new buyers requires a slightly different approach than you are used to with IT…

- Emergence of these new buyers is driven by business decisions and technical factors
  - Greater influence by new LOB buyers over technology
  - Self-service that provides independence from IT
- New buyers are more focused on *new systems*
  - Ease of development and time to market are essential
  - Experimentation is critical, requirements are not clear upfront
  - More focus on innovation to reach clients and optimize revenue
  - Self-service is required with less dependence on IT
- New buyers are *frustrated* with IT
  - Slow reaction times and too many bureaucratic processes
  - Lengthy lead times to deploy to production means missed market opportunities
  - Lengthy release cycles removes the agility to adjust to changing market conditions
  - Developers are unable to innovate quickly due to constraints of corporate standards
- IT feels pressured due to *competition with SaaS*
  - Needs to provide more automated, self-service environments
  - Concerned about losing control over governance and security

Refer to the Appendix for example conversations with these new buyers
Cloud Drives Better Business Economics

The number one reason to adopt cloud is NOT cost savings, it is **agility**. Cloud enables businesses to compete *faster*. And speed kills, a chess grandmaster would lose to a beginner that got four moves every turn.

**1. Cost Flexibility**
- Shifts fixed to variable cost
- Pay as and when needed
- CAPEX vs. OPEX

**2. Business Scalability**
- Provides limitless, cost-effective computing capacity to support growth

**3. Market Adaptability**
- Faster time to market
- Supports experimentation

**4. Masked Complexity**
- Expands product sophistication
- Simpler for customers and users

**5. Context-Driven Variability**
- User defined experiences
- Increases relevance

**6. Ecosystem Connectivity**
- New value nets potential new business

Source: IBV Analysis
Cloud Deployment Models

1. **Private Cloud** – A cloud dedicated for sole use and resides either on-premises or off-premises that is Client-managed or IBM managed.

2. **Public Cloud** – When the cloud is off-premises and resources may be shared between clients. Hundreds to thousands of clients with controls to enforce isolation between client’s data and resources. Most new ISVs are purely Public.

3. **Hybrid Cloud** – A combination of Public and Private clouds. Most enterprise clients are Hybrid.
SoftLayer Global Footprint

- 42 Data Centers WW
- 5 Continents
- 16 Countries
- Over $1.2B invested

Global DNS
IPv4/IPv6 dual stack
Global DDOS Mitigation
Global Internet Exchanges & Peering
Agenda

- **Cloud Solution Considerations**
- **Cloud Data Services**
  - Cloudant
  - dashDB
  - DB2 on Cloud
  - BigInsights on Cloud
  - Spark
  - DataWorks
  - Watson Analytics
- **CDS Integration Examples**
Why the Journey to Cloud-based Data Services?

MISSION

To provide the best experience for developers, data scientists and data engineers with a comprehensive set of rich, integrated cloud data services covering content, data and analytics.

FASTER INNOVATION

Instant provisioning saves weeks of data center setup

BETTER IT ECONOMICS

Pay as you go with no big up-front capital investments

LOWER RISK OF FAILURE

Fully managed 24x7 so you can focus on new development
Cloud Data Services is **Open For Data**

An **open portfolio** of self-service, composable data and analytic services for the developer, data science professional, and analytic architect. We help **transform** businesses and organizations to **build applications** and **gain new insights** better and faster.

<table>
<thead>
<tr>
<th>Comprehensive</th>
<th>Flexible</th>
<th>Trusted</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Brodest selection of data and analytic services available on multiple cloud platforms</td>
<td>• Open-sourced driven innovation</td>
<td>• Fully managed: 24 x 7</td>
</tr>
<tr>
<td>• Pre-built integrations across the portfolio</td>
<td>• Industry leading support for hybrid deployments</td>
<td>• Secure infrastructure</td>
</tr>
<tr>
<td>• Integrated with open data to gain deeper insights</td>
<td>• Bare metal, virtual, pay-as-you-go and reserved</td>
<td>• Mitigate risk and lower costs</td>
</tr>
</tbody>
</table>
Cloud Data Services Span a Variety of Needs

Customer Data Center (On-Premises)

Software
- Select, purchase, provision and operate hardware & software
- Complete control of every operating parameter

Appliance
- Expertly engineered hardware & software
- Customer responsible for operation
- Fixed configuration = less control

Cloud Image
- Expertly engineered cloud configurations
- IaaS manages hardware
- Customer has full control over software & operation

Managed Service
- IBM operates service and controls everything
- IBM delivers on SLA
- Customer concerns only with data

Cloud Data Center (Off Premise)

Control

Simplicity
Data on Premises – Workload Optimized Products

- Offerings for structured and semi-structured data, transactional and analytics workloads.
- Knit together with the InfoSphere IIG technologies for data movement and integration.

**BigInsights**
- Enterprise class Hadoop and real-time
- BigSQL for easier analytics
- IBM differentiators like GPFS

**Cloudant Local**
- Massively scalable
- Eventual consistency model
- Built for Systems of Engagement

**DB2 BLU**
- SQL interface
- ACID compliance
- Columnar, in-memory performance
- DB2 Built for Systems of Insight

**DB2**
- SQL interface
- ACID compliance
- Flexible HA and DR options
- Built for Systems of Record

**DataStage**
- On-Premises to Cloud ETL Software

**UNSTRUCTURED**

**STRUCTURED**
IBM Cloud Data Services

IBM CDS solutions combine workloads and data types for true hybrid services

**ANALYTICAL**

- **BigInsights on Cloud**
  - Spark for in-memory Hadoop
  - Built on IBM Open Platform
  - Bare metal performance
  - BigInsights enterprise features

- **Spark**

- **Cloudant**
  - Database as a Service (DBaaS)
  - Massively scalable for global data distribution
  - Eventually consistent data model
  - Built for mobile, Systems of Engagement

**UNSTRUCTURED**

**STRUCTURED**

- **Watson Analytics**
  - Analytics & Visualization Services

- **dashDB**
  - SQL interface, ACID compliant
  - Columnar, in-memory performance
  - BLU augmented with Netezza in-DB analytics
  - Built for Systems of Insight
  - MPP

- **dashDB Transactional**
  - SQL interface, ACID compliant
  - Optimized for transactional workloads
  - Built for Systems of Record
  - Oracle Compatibility

**TRANSACTIONAL**

- **DataWorks**
  - Data Refinery Services

- **DB2 on Cloud**
  - Hosted DB2 service
  - Flexible deployments for analytics or OLTP

**DashDB**

**BigInsights on Cloud**

**Cloudant**

**DashDB Transactional**

**DB2 on Cloud**

**Watson Analytics**

**DataWorks**

**IBM CDS solutions combine workloads and data types for true hybrid services**
Cloudant – Non-Relational Operational DBaaS

- **Distributed NoSQL “Data Layer”**
  - Powering Web, Mobile & IoT since 2009
  - Transactional JSON NoSQL Document database accessible using a RESTful API
  - Spreads data across data centers & devices for scalability & high availability

- **Available As**
  - A fully managed DBaaS
  - By you on-premises using Cloudant Local
  - Hybrid

- **Ideal for applications that require**
  - Massive, elastic scalability
  - High availability
  - Geo-location services
  - Full-text search
  - Occasionally connected users
  - Flexible database schemas that are fluid
Cloudant Is Fully Hybrid Cloud Enabled

- Public Cloud, Private Cloud or combine them into a Hybrid Cloud

<table>
<thead>
<tr>
<th>Cloudant Managed</th>
<th>Cloudant Local</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guaranteed Performance with Public Cloud DBaaS</td>
<td>The Power of DBaaS in the privacy of your data centers</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cloudant Dedicated</th>
<th>Cloudant Multi-tenant</th>
<th>Cloudant Local</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hosted &amp; Managed by Cloudant</td>
<td>Hosted &amp; Managed by Cloudant</td>
<td>Customer-hosted &amp; managed with Cloudant DevOps tooling</td>
</tr>
<tr>
<td>24x7 Premium Support</td>
<td>Community Support</td>
<td>24x7 Premium Support</td>
</tr>
<tr>
<td>Use for <strong>production deployment</strong>, development</td>
<td>Use for <strong>development</strong> &amp; prototyping</td>
<td>Use for production deployment, development</td>
</tr>
<tr>
<td>Monthly, per-node fee</td>
<td>Monthly, metered usage fee</td>
<td>Up-front perpetual license, or monthly, per-node fee</td>
</tr>
<tr>
<td>Available @ 30+ SoftLayer, AWS, Rackspace, Azure data centers</td>
<td>Available on SoftLayer, Rackspace, Amazon, Azure</td>
<td>On-premise or cloud platform of choice</td>
</tr>
<tr>
<td>Single-tenant clusters</td>
<td>Multi-tenant clusters</td>
<td>Single-tenant clusters</td>
</tr>
</tbody>
</table>
IBM dashDB – Analytics Warehouse as a Service

In-db analytics capabilities for best performance atop a fully-managed warehouse

- Fully-managed data warehouse on cloud
- **BLU Acceleration** columnar technology + **Netezza** in-db analytics
  - BLU in-memory processing, data skipping, actionable compression, parallel vector processing, "Load & Go" administration
  - Netezza predictive analytic algorithms
  - Fully integrated RStudio & R language
- Oracle compatibility
- Massively Parallel Processing (**MPP**)
dashDB Key Use Cases

Extend / Modernize
- Extend on-premises data warehouse environments to the cloud
- Flexible, cost-effective growth
- Hybrid cloud models support ground to cloud

Cloudant Analytics
- Easy synchronization of JSON to structured data
- Allows analytics via standard BI tools
- In-database predictive algorithms allow greater insight for Cloudant users than ever before

In-Database Analytics
- Robust predictive analytic algorithms
- Integrated with R
- Watson Analytics ready
- Analytics ecosystem with partners

Data Warehouse & Analytics Service
- Data warehousing and analytics in the cloud
- Cloud agility and flexibility
- Analytics for cloud data, data marts, and development & test environments
IBM Offerings – NEW dashDB TX (Initial GA April 12, 2016)

For apps that need:

- Elastic scalability
- High availability
- Data model flexibility
- Data mobility
- Text search
- Geospatial

Available as:

- Fully managed DBaaS
- On-premises private cloud
- Hybrid architecture

Excellent Transactional Performance

- Fully-managed transactional database as a service
- Row-organized tables for high transactional performance
- Oracle compatibility
- On disk data encryption and secure connectivity
- Two initial Enterprise plans
  - 2 cores, 8 GB memory, 500 GB SAN
  - 12 cores, 128 GB memory, 1.4 TB SSD
# A Sample of dashDB Data Load & Integration Options

## Systems of Record & Systems of Engagement

<table>
<thead>
<tr>
<th>IBM Cloudant</th>
<th>Oracle DB</th>
<th>IBM DB2</th>
<th>Load from Desktop (Excel, CSV) IOT + Web</th>
</tr>
</thead>
</table>

## Data Load & Refinement

<table>
<thead>
<tr>
<th>IBM DataWorks (Cloud Data Refinery)</th>
<th>IBM Datastage (On-Prem to Cloud ETL)</th>
<th>Aspera (High Speed File Transfer)</th>
<th>IBM Swift (SoftLayer Cloud Storage)</th>
<th>AWS S3 (Amazon Cloud Storage)</th>
</tr>
</thead>
</table>

## Reporting, Visualization, Exploration, and Analytics

<table>
<thead>
<tr>
<th>Tableau</th>
<th>SAS</th>
<th>Cognos</th>
<th>ESRI ArcGIS</th>
<th>SPSS</th>
<th>Aginity Workbench</th>
<th>Watson Analytics</th>
</tr>
</thead>
</table>

© 2015 IBM Corporation
DB2 on Cloud

- **Provides a Hosted DB2 Environment**
  - Provisioned via Bluemix
  - Hosted on IBM SoftLayer
    - Virtual private nodes (not shared) or bare metal depending on configuration size
  - Administered by your organization's administrators
    - e.g. OS updates, DB2 updates, backups, HA configuration
  - Paid for on a month-to-month basis via a subscription model with support included

- **Benefits**
  - Convenience without the loss of control on cost effective infrastructure
  - Four high performance hardware configurations and two database software tiers to match capability and affordability needs
  - BLU Acceleration
  - Native encryption support included in all configurations ensuring data remains secure in the cloud

- **Four T-shirt sized configurations**
  - Small, Medium, Large, X-Large and 2X-Large

- **Two versions of DB2 available**
  - IBM DB2 Workgroup Server Edition
  - IBM DB2 Advanced Enterprise Server Edition
Highlights of IBM's Database-as-a-Service Options

- DB2 on Cloud
- IBM Cloudant
- dashDB Transactional
- dashDB

Roll-Your-Own vs Hosted

Data, Admin, Software, Hardware
Compose Managed Platform of DBaaS Offerings

Compose is a cloud Managed Platform of Databases as a Service (DBaaS) that delivers a set self-service, production-ready Open Source services that can be adopted individually or as part of a fit-for-purpose data strategy

- **mongoDB**: Leader in the field of JSON document databases.
- **redis**: An open-source, blazingly fast, low maintenance, key/value store, often used for data caching.
- **elasticsearch**: Combines the power of a full text search engine with the indexing strengths of a JSON document database.
- **PostgreSQL**: A powerful, open source object-relational database that is highly customizable.
- **RethinkDB**: A JSON document based, distributed database with an integrated administration and exploration console.
- **RabbitMQ**: Asynchronously handles the messages between your applications and databases, allowing you to ensure separation of the data and application layers.
- **etcd**: This key/value store holds the always-correct data you need to coordinate and manage your distributed applications server clusters.
IBM Compose – Services & Deployments

Compose services are deployable to both SoftLayer and AWS

Available on SoftLayer as:
- IBM-Managed service
- Public Multi-Tenant
** Self-Hosted services coming soon

Available on AWS as:
- IBM-Managed service
- Self-Hosted service
- Public Multi-Tenant

IBM SoftLayer

Amazon AWS
IBM BigInsights on Cloud
Making the Cloud work for you

**Build**
- Ready-to-run, fully managed Hadoop clusters in the cloud
- IBM Open Platform - 100% open source Hadoop; will align with ODP
- Based on proven, performant reference architectures

**Manage**
- Key platform components monitored for availability
- Hadoop, OS and BigInsights patched and maintained
- Ambari cluster manager for complete control

**Support**
- 24x7 cloud operations and support team
- Access to deep Hadoop expertise
- Faster time to problem resolution

**Protect**
- Deployed in world-class, secure SoftLayer data centers
- Dedicated physical machines with HDFS encryption*
- Firewall, Knox, Kerberos, LDAP secure the clusters
IBM BigInsights for Apache Hadoop

Consume BigInsights based on Analytic Workload and Skillsets available

IBM BigInsights Data Scientist
- Text Analytics
- Machine Learning on Big R
- Big R
- Big SQL
- BigSheets

IBM BigInsights Enterprise Management
- POSIX Distributed File system
- Multi-workload, Multi-tenant scheduling

Small Nodes
- Basic data extraction, transformation, file processing, search
- 16 cores, 64GB RAM, 20 TB data disk, 8 TB OS disk, 10 GB dual path network

Medium Nodes
- Data warehouse optimization – store new data or extend warehouse
- 16 cores, 128GB RAM, 28 TB data disk, 8 TB OS disk, 10 GB dual path network

Large Nodes
- Advanced Analytics – intensive data processing
- 16 cores, 192GB RAM, 32 TB data disk, 8 TB OS disk, 10 GB dual path network

IBM Open Platform with Apache Hadoop
What is Spark?

Spark is an open source in-memory application framework for distributed data processing and iterative analysis on massive data volumes.

- An Apache Foundation open source project; not a product
  - Spark is open, accelerating community innovation
- An in-memory compute engine that works with data; not a data store
  - Spark is fast—100x faster than Hadoop MapReduce
- Enables highly iterative analysis on large volumes of data at scale
  - Spark is about all data for large-scale data processing
- Unified environment for data scientists, developers and data engineers
  - Spark supports agile data science to iterate rapidly
- Radically simplifies the process of developing intelligent apps fueled by data
  - Spark can be easily integrated with IBM solutions
IBM Analytics for Apache Spark – Spark-as-a-Service

- Access to Spark’s next-generation performance and capabilities, including built-in machine learning and other libraries
- Pay-as-you-go – Pay only for what you use
- No Vendor lock-in – 100% standard Spark that runs on any standard distribution
- Elastic Scaling – Start with experimentation, extend to development and scale to production, all within the same environment
- Quick start – The service is immediately ready for analysis, skipping setup hurdles, hassles and time
- Peace of mind – fully managed and secured, no administration necessary
Spark Adoption…

IBM is adopting Spark throughout the IBM Analytics portfolio

---

**Within Platforms**

- **IBM Open Platform**
  Spark included with distribution

- **BigInsights for Apache Hadoop**
  Spark included with IOP and BigInsights runs on IOP

- **IBM Streams**
  Spark is a data-at-rest engine

- **IBM DataWorks**
  Spark is the distributed data processing engine

---

**Within Solutions**

- **IBM Commerce**
  
  - **DemandTec**
    Spark used for marketing automation
  
  - **Unica**
    Spark used for pricing automation

- **Watson Analytics**
  Spark used to analyze and refine data

- **Watson Health**
  Spark used to analyze massive amounts of personal health data

---

**Future**

- **IBM SPSS**
  IBM PureData for Analytics
  Enable Fluid Query to query Spark data

- **IBM Research**
  Currently has over 30 active projects utilizing Spark
  …and many more
IBM® DataWorks™ Forge

An intuitive application and user experience to find, shape, enrich and deliver trusted data with confidence

- Empowers users to find relevant and useful data
- Automatically profiles and classifies data on ingestion
- Delivers profiling and quality metrics to visualize and understand the data
- Easily shape and enrich the quality of data by filtering, joining, sorting and more...
- Allows users to save and run activities and deliver the data to their chosen targets
IBM Watson Analytics

Self-service analytics capabilities in the cloud

Single Analytics Experience

Fully Automated Intelligence

Natural Language Dialogue

Guided Analytic Discovery
IBM Watson Analytics

Single Interface … Refine > Explore > Predict > Assemble
IBM Watson Analytics

Explore

Continue to ask questions about your data using natural language.

Click on the new insights provided by Watson Analytics and explore the text and visual results.
IBM Watson Analytics

Predict

Additional insights discovered for you

Identified predictors of your target analysis

Pin your results for use in dashboards and storytelling

Automated data visualizations
Agenda

- **Cloud Solution Considerations**
- **Cloud Data Services**
  - Cloudant
  - dashDB
  - DB2 on Cloud
  - Compose
  - BigInsights on Cloud
  - Spark
  - DataWorks Forge
  - Watson Analytics
- **CDS Integration Examples**
A Car Manufacturer
requires a data management or analytic solution to understand campaign effectiveness.
Use Case: Car Manufacturer

Company Background
– A multinational car manufacturer is equipping their 2015 and later models with an "interactive" infotainment system.

Need
– A data management or analytic solution to understand campaign effectiveness.

Success Criteria
– Take unstructured data, automatically format it, and report against it in the cloud, allowing the company’s marketing team & merchants to segment customers and measure campaign effectiveness.

Solution & Results
– Provisioned a prototype environment of dashDB with Cloudant in 24 hours illustrating the speed and agility possible
– Flexibility and power of the IBM portfolio was critical for the project’s success
– The car manufacturer is getting a lot of value and analytics from their infotainment systems.
Use Case: Car Manufacturer– Architecture Details

Vehicle Data
- Bulk load of 6-8 million accounts, 28 millions cars (~1TB of data)
- Deltas continued to be load for user/vehicle data

Other Data Sources

IBM Cloudant
- Softlayer Dallas
- 3 node production cluster
- 3 node dev cluster

IBM dashDB
- Softlayer Dallas
- 4 TB production system
- 4 TB dev system

IBM Datastage
- Softlayer Dallas
- Data movement into SFDC and into dashDB

JSON data to relational
Cloudant Schema Discovery Protocol

IBM Cognos

salesforce
Client Example High Level Architectural

IBM Data Centre (SoftLayer)

VLAN1

SPSS
InfoSphere (ETL)
SFTP Server

10GB Adapters

DashDB Instance 1
DashDB
DashDB Instance 2

VLAN2

IBM Data Centre (SoftLayer)

Source DB
Source DB
Other..

Customer Data Center

Client Data Sources

VPN Access 1Gbps Bandwidth

1Gbps Bandwidth
Second Client Example
Case Study: Retail business derives insights with Cloud Analytics

Beacons in shopping zones stream customer location. Kiosks provide cognitive interaction.

IoT Data

Weather Company

Ingest bulk Data

Ingest Streaming Data

Ingest On-prem Data

Data Lake and Analytics in the Cloud

Data Stores

Governance

Analytics Processing

Scenarios

Explore & Visualization
- Analyze declining sales and discover causes

Advanced Analytics
- Analyze data in context, build model for cross-sell campaign

Data Integration
- Build data integration jobs

Application Development
- Build applications for users

Business Analyst

Data Scientist

Data Engineer

App Developer

Retail store

Outdoor

Mountain

Golf

Camping

Persona

Outdoor

Mountain

Golf

Camping

Customer Marketing Sales History
Details: End-to-End Analytics with IBM DataWorks Platform

Cloud services ingest all types of data

‘Fit for Purpose’ stores and governance for Enterprise Data Lake

‘Spark as a Service’ enables scalable, high performance analytics on all cloud data

Cloud tooling enables all user personas

IBM DataWorks platform
Introduction: Meet the Business

“The Great Outdoors”

The Great Outdoors is a national chain that carries sporting goods including golf, camping and outdoor protection. In addition to their brick and mortar locations, they also sell merchandise online.

Business Problem: Declining sales in outdoor protection
Sales declining year over year

Gross Margin indicates room for a *pricing* play
Demo: Increase Sales with Real-Time Targeted Offers

Ask Watson
Describe what you are looking for

2 man tent
THANK YOU
Questions?