New Modeling Challenges: Big Data, Hadoop, Cloud

Karen López
@datachick
www.datamodel.com

Karen Lopez

Love Your Data

InfoAdvisors.com
@datchick

Senior Project Manager & Architect
Disclosure

I’m a Data Architect / Modeler

I am biased

You have been warned

POLL: Who Are You?
POLL: Architect Much?

POLL: Cloud Much?
What we will be doing

- Terminology, Foundations and Thoughts.
- Modern Architectural Components
- Architecting versus building
- Cost, benefit, risk
- Demos!
What would you like to do?

Outcomes

- Concepts
- Features
- Existing Tech
- New Tech
- Hybrid
- Evaluate
- Pilot
- Production
- Benefits
- Trade-offs
- Architects
- Devs
- DBAs
- Modeling
- Dev
- Test
- Costs
- Trade-offs

What
Where
When
Why
Who
How
How much
Terminology and Thoughts

Because we need a vocabulary to share ideas

- Hadoop
- BIG DATA
- Unstructured data
- Cloud
- Analytics
- Variable Data
- Logical Data Model
- Data Lake
- Data Reservoir
- Polyglot
- Volume
- Write-Optimized
- Read-Optimized
- Scale Up / Out
- Data versus Processing
- Schemaless

© 2015 InfAdvisors
Polyschematic Relational Read-Optimized Persistence

HA / DR

On Premises

Velocity

Volume

Business Intelligence

External Data

Physical Data Model

JBIC data

Data Swamp

ROI

Architect

Business

Data Protection

Technology

Cost, Benefit, Risk

Fit
“BIG DATA”

[x] Vs

“Data so big it’s awkward to work with”

Always capitalized Big Data

A confusing term because it defines what it IS NOT.

“NoSQL”

Scale

Not SQL?

Not Relational?

Not Only Relational?

A confusing term because it defines what it IS NOT.
**Terminology**

<table>
<thead>
<tr>
<th>ACID</th>
<th>Atomic, Consistent, Isolated, Durable</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASE</td>
<td>Basically available, Soft state, Eventual consistency</td>
</tr>
</tbody>
</table>

**Eventual consistency**

**Schemaless**

Constraints / Have-to/ MUST / OBEY / Rigid / Inflexible

---

**Relational**

- On purpose
- With many benefits
- Write-optimized
- This is a FEATURE
- Data integrity
- Data quality
- Consistent

- Tables with rows
- same columns with the same datatypes
- with the same constraints
- with the same domains

- Trans- action- optimized
Data Modeling Now

- What Data Models?
- Reverse engineered data models (diagrams)
- Physical Data Models
- Faux Logical Data Models
- Logical Data Models
- Conceptual Data Models

Data Models - Traditional Process

- Conceptual (Data) Model
- Logical Data Model
- Physical Data Model(s)
Traditional Data Architect Involvement

The Big Data Story

Lots of data
Coming at us fast
Lots of variety in format & quality
We want all the data
Highly available
“It’s web scale”
What do we really mean by scale?

- Bringing computing to the data
- Massively parallel processing
- Cheap, commodity hardware, but lots of it
- Optimized for Query/Reads/Questions/Telling stories

Can we fit another buzzword in?

Cloud

- Enable on-demand scaling
- Pay as you go pricing
- Click to deploy
- Service licensing, not product licensing, if any
- Managed by others, not your data center
But The Cloud is Different

- Fine Tuning...nope
- Patching, maintaining
- Skills
- Professional development
- Optimized subsystems
- Putting off upgrades
- How you work
- Who you work for
- How you get new features
- How you stay up to date
- How you think about problems

We’ve been down this road before...

- Traditional transactional applications
- Reporting-optimized tables/structures
- Data Warehouse / Dimensional Modeling
There was a lot of contention...

NoSQL, Not Only SQL

- Relational
- Graph
- Columnar/Column Family
- Key Value
- Document Databases
- Others
Graph Databases

Key Value Pair
Columnar

Existing Table → Row Groups → Column Segments → Columnsstore

Compacted column segments are added to the columnsstore.

Modern Architectural Components & Concepts
Hadoop Ecosystem (a Zoo)

- **Distributed Storage**: HDFS
- **Query**: Hive
- **Distributed Processing**: MapReduce
- **Scripting**: Pig
- **Metadata**: HCatalog
- **Query**: HiveQL
- **SQL-like Query**: Impala
- **Distributed Processing**: MapReduce
- **Resource Scheduling**: YARN
- **Distributed Storage**: HDFS
- **Event Pipeline**: Flume
- **Pipeline/Workflow**: Oozie

Legend:
- Red = Core Hadoop
- Blue = Data processing
- Purple = Microsoft integration points and value adds
- Yellow = Data Movement
- Green = Packages

© 2015 InfAdvisors
MapReduce

- Shuffle and sort data
- Get from large data to smaller data
- Parallel processing

Hive

- SQL-like query language
- Abstraction on top of MapReduce
- Metastructure on top of HDFS
HBase

Column Family NoSQL database
- Key Value Store
- Hundreds of Millions/Billions of rows

Based on Google Big Table

Another type of HDInsight Cluster

Large, Schemaless (really, Schema on read)
- Optimized for retrieving specific rows from large datasets

Strictly Consistent

Classic DW Architecture

OLTP DB

OLTP DB

External Data

ETL

Staging/ETL DB

EDW

Data Mart

Data Mart

Data Mart

On Premises
Modern DW Architecture

Hadoop
- Distributed Processing (MapReduce)
- Distributed Storage (HDFS)
- Distributed Storage (Blob Storage)

ETL

Cloud And/Or On Prem

Data Mart
Analytics Mart

What’s new

Hadoop
- Distributed Processing (MapReduce)
- Distributed Storage (HDFS)
- Distributed Storage (Blob Storage)

Storage

External Data

Icing ☺

Computing

© 2015 InfAdvisors
What’s new

Analytics Mart

Scale and Nodes

Scale up versus scale out
Nodes, not beefier servers
Data stays, clusters come and go
Quick Create for playing
Custom Create for architecting a solution
“Every design decision should include cost, benefit and risk”

- Karen Lopez
Cloud vs On-Premises

On-Premises
- Large long term implementations
- Well known workloads
- Shared clusters
- Large initial investment

Cloud vs On-Premises
- Short Term Use
- Rapid Scale
- Test Use Cases
- Pay as you go
- Internet data source

Windows vs. Other OSs (Linux)
Graph Databases

There’s a book

And it’s FREE!
Graphdatabases.com
Document Databases

JSON, BSON, XML, YAML...

A second document might be:

```json
{
    FirstName: "Jonathan",
    Address: "15 Wanamassa Point Road",
    Children: [
        {Name: "Michael", Age: 10},
        {Name: "Jennifer", Age: 8},
        {Name: "Samantha", Age: 5},
        {Name: "Elena", Age: 2}
    ]
}
```

Key Value Pair

![Key Value Pair Diagram](image-url)
### Key Value Pair

![Diagram of blog keyspace]

### Cassandra CQL

**SELECT** balance FROM accounts WHERE account_id = 3476

**CREATE TABLE monkeySpecies (**
  - species text PRIMARY KEY,
  - common_name text,
  - population varint,
  - average_size int
) WITH comment='Important biological records'  
AND read_repair_chance = 1.0;
\[ \text{CREATE TABLE Retail.Product_Dimension (} \] 
\text{Product_Key integer NOT NULL,} 
\text{Product_Description varchar(128),} 
\text{SKU_Number char(32) NOT NULL,} 
\text{Category_Description char(32),} 
\text{Department_Description char(32),} 
\text{Package_Type_Description char(32),} 
\text{Package_Size char(32),} 
\text{Fat_Content integer,} 
\text{Diet_Type char(32),} 
\text{Weight integer,} 
\text{Weight_Units_of_Measure char(32),} 
\text{Shelf_Width integer,} 
\text{Shelf_Height integer,} 
\text{Shelf_Depth integer \}); } \] 

\[ \text{SELECT balance FROM accounts WHERE account_id=3476} \]
The CLOUD

Infrastructure as a Service (IaaS)
Platform as a Service (PaaS)
Database as a Service (DBaaS)

Let’s talk billing....

Per hour
Per node
  head node
  security node
  data nodes
Data egress, not ingress
Geo-redundancy
Where Data Models Can Help

Create a Physical Data Model of Key Data Sources
- One Entity per tab or Worksheet
- Someone Normalized (master data + combinations)

Model the metadata
- definitions
- gotchas
- expected domains
- metadata

Published in a shareable portal
- Cross model searchability
- Data Lineage
- Resuse

The Data Story

IRS Energy Tax Credit Data*


Retail Data Warehouse

Power Query & Power View

HDInsight

*http://www.dsireusa.org/incentives/incentive.cfm?Incentive_Code=US37F
Discussion - Who

- What technical skills are needed?
- What professional skills are needed?
- What personal characteristics are needed?

Discussion - Tools

- What data tools are needed?
- What data architecture tools are needed?
- What project management tools are needed?
Discussion - Money

- How much does a traditional DW/BI project cost?
- How much does operations & maintenance cost?
- How much does the software cost?
- How much do servers cost?

Modern Data Architect Involvement

- Project Initiation
- Architecture and Infrastructure Design
- SW + Data Requirements
- Development
- Deployment
Anti-patterns for new DBs & Data

“Schemaless” means faster delivery

It doesn’t matter if data changes
We don’t need a data architect
We don’t need ANY architects

We don’t have to steward data

We don’t have to understand data

Its self-documented

Cc: Kazuhisa OTSUBO

© 2015 InfAdvisors

The home of the U.S. Government’s open data
Here you will find data, tools, and resources to conduct research, develop web and mobile applications, design data visualizations, and more.

GET STARTED
SEARCH DATA PROJECTS

BROWSE TOPICS

data.nasa.gov
What type of data are you looking for?
Galaxy Evolution Explorer (GALEX)
From its launch into low-Earth orbit on April 28, 2003 the GALEX/IAN camera operated almost continuously. The Project, sponsored by the American Space Agency and hosted in space, is one of a

NASA Standards
The NASA Technical Standards Provide access to NASA standards standards are

canada.gc.ca

English Indices of Deprivation 2010
The English Indices of Deprivation 2010 presents a range of measures of deprivation at local and national levels in England, using data from local authorities to national archives and views.
10 Tips for Architects

1. Understand the use cases for hybrid technologies.
2. Evaluate / profile your data requirements for suitability for each database/datastore type.
3. Understand the licensing / editions for commercial database features / products.
4. Hadoop and other NoSQL technologies are optimized for read.
5. Data Modeling and Analysis still happens, just later*.
10 Tips for Architects

6. Hands on experience makes a real difference in your understanding.
7. Evaluate which analytical features each product supports.
8. Test your current development tools for support.
9. Test your database design / data modeling tools.
10. Leverage your existing metadata / models.

Finally...

There business models of Open Source technologies means there are hundreds, maybe thousands of new things to learn. That’s both a pro and a con for the architect.
Outcomes

- Concepts
- Features

- Existing Tech
- New Tech
- Hybrid

- Evaluate
- Pilot
- Production

- Benefits
- Trade-offs

What | Where | When | Why
---|---|---|---
Concepts | Existing Tech | Evaluate | Benefits
Features | New Tech | Pilot | Trade-offs
Hybrid | Hybrid | Production | Benefits

- Architects
- Devs
- DBAs

- Modeling
- Dev
- Test

- Costs
- Trade-offs

Who | How | How much
---|---|---
Architects | Modeling | Costs
Dev | Dev | Trade-offs
DBAs | Test | Benefits

Wrap up

- NoSQL = SQL + NoSQL

External Data will become more important

Modern Data Architecture = Hybrids, not either/or
Thank You

I’d love to help you with your big data, nosql, cloud or other projects.
karenlopez@infoadvisors.com