Continuous analytics: Integrating the data hub in a devops pipeline

Arturo Bayo (Synergic Partners)
Álvaro Fernández (Santander Spain)
Data is driving today’s digital business strategy

DATA CENTRIC STRATEGY

Data is the new business currency

Data growth means new opportunities for every organization

Companies seek out actionable insights via connecting data
The data hub is the core in delivering a successful data strategy.

Organizations struggle to find the single version of truth among their data sources.

40% of data-driven organizations in the US will combine all data into a single repository.

Source: Forrester Consulting

This strategy has profound security and governance implications to ensure privacy regulations compliance.

#StrataData
### Design considerations for the data hub architecture

<table>
<thead>
<tr>
<th><strong>STORAGE</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DATA GOVERNANCE</strong></td>
</tr>
<tr>
<td><strong>PROCESSING</strong></td>
</tr>
<tr>
<td><strong>SECURITY</strong></td>
</tr>
</tbody>
</table>
Design considerations for the data hub architecture

STORAGE

1. Using heterogeneous storage on datanodes

2. Configuring HDFS storage policies

- **Data**
  - Live
  - History

- **Storage Types**
  - SSD: SSD
  - DISK: HDD (SAS)
  - ARCHIVE: HDD (SATA)

ALL_SSD
COLD
Design considerations for the data hub architecture

DATA GOVERNANCE

1. Defining a fair filesystem structure for the data lifecycle

2. Using an information catalog to understand data proliferation across the hub

- Cloudera Navigator
- Apache Atlas
Design considerations for the data hub architecture

1. Designing an efficient YARN configuration for the scheduler
   - Leverage dynamic resource pools if using FairScheduler (e.g. CDH)
   - Configure cluster partitions if using CapacityScheduler (e.g. HDP)

2. Calculating fine-grained resource isolation for critical services
   - CPU
   - Memory
   - I/O weight

Production

- NM1
- NM2
- NM3
- NM4
- NM5
- NMn

Processing

- YARN (50%)
- IMPALA (30%)
- HBASE (20%)

#StrataData
Design considerations for the data hub architecture

SECURITY

1. Integrating the data hub with a directory service and building a clean tree structure

2. Delegating authorization on services such as Sentry or Ranger. Don’t set ACLs manually!

3. Installing data encryption services only if needed, and chose encryption zones carefully if so

Apache Sentry RBAC

Data access control managed by the directory service
Putting it all together: data hub reference architecture

- **Ingestion**
  - Batch
  - ftp/nfs
- **Real time**
  - Near real time
- **Queueing**
  - Storm
- **Integration**
- **Processing**
  - Real time (RT)
  - Batch (BATCH)
- **Production**
- **Analysis**
  - Spark
  - Hive
- **Data Governance**
  - Metadata
  - Data Flow
  - Data Security
- **Storage**
- **Monitoring**
- **Exportation**
  - Active Directory
  - OpenLDAP
- **User roles**
  - Export
  - Use cases
  - Exploration

#StrataData
Architecture alone is not enough

Implementing the data hub is just the starting point

The real goal is building products which provide business value
Big data deployment stage worldwide

15% Knowledge Gathering
14% Developing Strategy
30% Piloting and Experimenting
15% Deployed

73% Already invested
OR
Plan to invest

Source: Gartner
Devops pipeline for analytics projects
Continuous Analytics Pipeline

Container platform

Data Hub architecture

Source Code

Build

Test

Artifact Repository

DEV ENVIRONMENT

Test

Promote

Test

Deploy

Config

DATA HUB

Integration

Production
Continuous Analytics Pipeline: implementation
Analytical use case at Banco Santander
A different approach to customer segmentation

Huge increase of available information to detect customer relationships

Democratization of A.I. and machine learning algorithms

The Social Customer
Customers are highly connected

- Married two kids
- My parents
- Owns a pharmacy
- My friends
- Employees
- Pays rent to landlord

Ana
A central customer

Distance = 1
1,677 nodes
(44% active customers)

1,800 employees
30 subsidiary companies

Payments: 532 companies

Distance = 2
54k nodes

Real network

#StrataData
Product propagation in the network

Exponential propagation in communities like schools, where word of mouth was the main driver of acquisition
Product propagation and node profile

How node attributes help to propagate a message in the network

<table>
<thead>
<tr>
<th>Women</th>
<th>Men</th>
</tr>
</thead>
<tbody>
<tr>
<td>28%</td>
<td>17%</td>
</tr>
</tbody>
</table>

Node Attributes:

- Doctors: 53%
- Commercial: 40%
- Hairdressers: 35%
- Taxi drivers: 30%
- Agriculture: 15%
- Truck drivers: 13%
- Police: 8%
- Family: 40%
- Work: 27%
- None: 18%
Segment using the social value of customers

Customer centrality as key metric

Individual value of a customer

Social value same customer
The Big Chase*

Using the network to find the optimal path to acquire a new customer

It’s 6 times easier to acquire a new customer if it’s already connected to my customers

* Title of the article published in the scientific journal “Decision Support Systems”, as a collaboration with Universidad Carlos III
A new way of customer management

01 Segment customers according to its social value

02 Better target potential customers

03 Identify high churn risk customers

04 Improve risk profiling

05 Make better commercial decisions

#StrataData
To wrap up: Key learnings

1. **Cross-company initiative**: The data hub strategy will not succeed without strong collaboration.
2. **Data sharing**: The architecture must support data sharing in a governed and secure way.
3. **Business milestones**: Be aware that data architecture is crucial, but not the final goal.
4. **Production focus**: Pay attention to application lifecycle processes.
5. **Value through AI**: Focus on getting fast insights from data through analytics.
6. **Agile industrialization**: Bring devops and continuous analytics culture to the organization.
Continuous analytics: Integrating the data hub in a devops pipeline

Arturo Bayo
abayo@synergicpartners.com | @arturobayo

Álvaro Fernández
alvarfernandez@gruposantander.com