Hadoop and Spark at ING

An overview of the architecture, security, and business cases at a large international bank

Bas Geerdink

09/28/2016
Bas Geerdink

IT Manager at ING Analytics

Academic background: AI, Informatics

Working in IT since 2004

At ING since 2013

@bgeerdink

https://nl.linkedin.com/in/geerdink
# Agenda

1. **Strategy**  
   Why ING is a data-driven enterprise

2. **Data**  
   What we store and analyze

3. **Use cases**  
   What we do with Hadoop and Spark

4. **Architecture & Technology**  
   How we implement the tools

5. **Conclusions**  
   Lessons learned and the road ahead
Strategy
ING is a top financial enterprise, operating since 1881

- Customers: 33 Million Private, Corporate and Institutional Customers
- Countries: 41 In Europe, Asia, Australia, North and South America
- Employees: 52,000

Map showing distribution of countries and markets:
- Market leaders Benelux
- Challengers
- Growth markets
- Commercial Banking
ING’s Think Forward strategy requires a data-driven approach

**Purpose**

EMPOWERING PEOPLE TO STAY A STEP AHEAD IN LIFE AND IN BUSINESS

**Customer Promise**

CLEAR AND EASY  ANYTIME, ANYWHERE  EMPOWER  KEEP GETTING BETTER

**Strategic Priorities**

CREATING A DIFFERENTIATING CUSTOMER EXPERIENCE

1. EARN THE PRIMARY RELATIONSHIP
2. DEVELOP ANALYTICS SKILLS TO UNDERSTAND OUR CUSTOMERS BETTER
3. INCREASE THE PACE OF INNOVATION TO SERVE CHANGING CUSTOMER NEEDS
4. THINK BEYOND TRADITIONAL BANKING TO DEVELOP NEW SERVICES AND BUSINESS MODELS

**Enablers**

SIMPULATE & STREAMLINE  OPERATIONAL EXCELLENCE  PERFORMANCE CULTURE  LENDING CAPABILITIES
We are committed to helping our customers stay a step ahead by making their data work harder for them.
Data Everywhere
There is a lot of data in the bank, in many sizes and shapes.

Most reference data is not stand-alone, but belongs to one of the other data categories.

1. PARTY DATA
   - Name, address, place of birth, statutory address, ...

2. AGREEMENT DATA
   - Terms, conditions, ...

3. TRANSACTION DATA
   - Payments made, interest accrued, ...
   - Country codes, currency codes, ...

4. REFERENCE DATA
   - System down times, processing throughput times, ...

5. PRICING DATA
   - Exchange rates, stock prices, ...

6. OPERATIONAL, SYSTEM, PROCESS DATA

DERIVED DATA BY FINANCE

DERIVED DATA BY RISK

DERIVED DATA BY HR

DERIVED DATA BY ..
There are three types of data streams at ING

<table>
<thead>
<tr>
<th>Type</th>
<th>Velocity</th>
<th>Volume per update</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Batch files</td>
<td>Hours / days</td>
<td>~ 1 GB</td>
<td>Mortgage balance</td>
</tr>
<tr>
<td>Increments</td>
<td>Seconds</td>
<td>~ 100 kB</td>
<td>Address change</td>
</tr>
<tr>
<td>Real-time events</td>
<td>Milliseconds</td>
<td>~ 100 B</td>
<td>Clickdata, transactions, stock prices</td>
</tr>
</tbody>
</table>

Is a batch a special form of streaming data?

Is a stream a mini-batch?
We combine batch and real-time data to give customers insight
Use Cases
How do we want to use data at ING?

“On Target” = the right message at the right time in the right channel to the right customer

What does ‘right’ mean?
- We want to be as relevant as possible
- ‘right’ = interesting enough to react positively (i.e. to click)
Use case #1: Real-time Account Forecasting

- Predict transactions and future balance based on historic transactions
- Fast data flow, calculations within 200ms
Use case #2: Outbound Communications

- Personal messages, based on large data volumes, smart segmentation and business rules
- Omnichannel: mobile, internet, letters, email, phone
Use case #3: Personal, Relevant Financial Advise

- Currently, real-time analytics determine the Next Best Action for a customer
- We'll extend this to actionable insights that help customers become financially fit
- From rigid customer segments to behaviour-driven segmentation
Architecture & Technology
All Analytics applications follow the same pattern
The Data Lake is the “memory” of the Bank holding all data relevant for reporting, advanced analytical and data exchange.
We use four different Hadoop usage patterns
The technology of the data lake depends on the data type, access requirements and enterprise-readyness.
Hadoop is integrated with Corporate Directory for secure role-based access
Architecture #1: Real-time Account Forecasting
Architecture #2: Spark 2.0 Dataset / SQL batch processing for Outbound Communications
Architecture #3: Actionable Insights Microservices

Static Data -> Feature Extractor -> Features -> Learning Environment

Event Data -> Streaming Analytics Engine

Batch 2 Stream -> Scoring API

ML models

Get Features -> Business Rules -> Score Insights -> Insights API -> Insights Front-end

Feature Extractor

Static Data

Features

Learning Environment

Streaming Analytics Engine

Insights API

Insights Front-end

ML models

Get Features

Business Rules

Score Insights

Insights API

Insights Front-end
## Challenges of using Hadoop and Spark in the enterprise ... and how to overcome them

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Security</td>
<td>Implement security from day 1. Tools: Ambari, Kerberos &amp; Ranger, improved by ING</td>
</tr>
<tr>
<td>Data Quality</td>
<td>Data owners, governance, metadata management, quality checks at the source</td>
</tr>
<tr>
<td>Zero-downtime patches and releases</td>
<td>Full DTAP, automated installs, no GUIs, CD pipeline: Git, Jenkins, Ansible, CA Lisa</td>
</tr>
<tr>
<td>Open Source</td>
<td>Code checks, Hortonworks distribution, active community involvement</td>
</tr>
<tr>
<td>Many data sources</td>
<td>Automate ETL, internal knowledge sharing (communities and chapters)</td>
</tr>
<tr>
<td>Many different specialties</td>
<td>Work agile in multidisciplinary teams (squads): engineers, data scientists, customer journey experts</td>
</tr>
</tbody>
</table>
Why is Hadoop the technology of choice for ING, moving forward?

- Bigger data volumes (static & streaming) to handle
- Enterprise-readyness and security of the ecosystem is continuously improving
- We now have 2 years of experience in Hadoop, with several teams working across Europe
- With Spark we have a tool that is very well suited for multidisciplinary teams, where data scientists work on the same user stories as engineers
Questions
Have a nice future!
(we’re hiring!)
Follow us to stay a step ahead

ING.com

@ING_News
Facebook.com/ING

LinkedIn.com/company/ING
YouTube.com/ING

SlideShare.net/ING
Flickr.com/INGGroup