Big Data Analytics in Telecommunication

Nokia NET Technologies & Innovation / Norbert Kraft

code::dive conference / Wroclaw / 05-Nov-2015
Agenda

Intro

Telecommunication Data & Use Cases

Anomaly & Root Cause Detection

Time Series Analysis & Prediction

Software & Methods

What we do ...

Why we do ...

How we do ...
Short Introduction

- Software researcher & data analyst
- Nokia Technology & Innovation
- Long history in SW development
- Project Leader NDI research project:

‘Network Data Intelligence’
What is a data analyst?

A Person ...

who knows more about programming than a mathematician ...

... and who knows more about statistics than a programmer.
Research Project

**Network Data Intelligence**

- Nokia research project
- Technology exploration
- Generate new insights in telecom data
- Raise new business opportunities
End to End Mobile Broadband
More Than an End Device

Dock + O/E conversion
Mini BTS
Standalone GPS module
External directional antenna

User Entity
BTS/eNodeB
MME
HSS
PCRF
Internet

User Data
Signaling

© Nokia Networks 2015
Some (estimated) Numbers ...

German Telekom (2012)

- 36.6 Million Subscribers for German Telekom
- Total of 113 Million Subscribers in Germany
- ~70 000 Radio Cells in Germany
- ~100 Million GBytes traffic volume (*2011)
- xxx.xxx.xxx.xxx Number of Calls & SMS per Day
- xxx.xxx.xxx.xxx Number of Internet connections
- SmartPhone is always ‘ON’

Source: Bundesnetzagentur from 2012
Total number of Radio Cells: Munich Example
Some Secret about Big Data (... you might have never heard)  
Big data is useless ... only Information counts

<table>
<thead>
<tr>
<th>Id</th>
<th>IMSI</th>
<th>IMEI</th>
<th>Radio cell id</th>
<th>RNC</th>
<th>Error code</th>
<th>IP address</th>
<th>ports</th>
<th>...</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>xxx</td>
<td>2345</td>
<td></td>
<td>45</td>
<td>1</td>
<td>0x0af</td>
<td>10.1.</td>
<td>80</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Customer A has some trouble, specifically if he enters radio cell B during business hours, because he is using an old fashioned smart phone with a wrong setup and additionally his office location has a bad radio coverage.
Agenda

- Intro
- Telecommunication Data & Use Cases
- Anomaly & Root Cause Detection
- Time Series Analysis & Prediction
- Software & Methods
What the Operator (needs to ...) know about ...
Mobile Network Data on ‘Signaling’ not on Content!

**User**
- User Identity
  - IMSI
  - A/B No
  - MSIDN
- Location
  - Cell location
  - Higher precision with triangulation on signal strength
- Service Usage
  - URL
  - IP ports
  - User Agent
  - Device OS
- Tariff
  - Personal Data
  - Revenue
  - Sex, Address
- Data Volumes
  - Bytes up/down
  - Call/SMS Length

**Network**
- Network Element Status
- Configuration Data
- Performance values
- Alarms
- SW Logs & Traces
- CDRs
- Highly structured

**Structured vs. Unstructured**

© Nokia Networks 2015
Attention

Network Data is **Personal Data**

!!!!

**Meta data**
- Network operators have the right to use this data for management purposes
  - Billing
  - Fault diagnosis
  - Network improvement
  - Support activities

**User data**
- Strictly limited access allowed with judicial order acc. local laws

• Strictly limited by (inter)national laws
• Very complex field under continuous change
• Different views in different countries
• Restrictions on use beyond network management scope
• Usage requires customer permission
Use Cases from the NDI Project

Subscriber Relations

Subscriber Mobility

Real Time Service Dash Board

Radio Cell Performance

Data Mining

Big Data

Mobile Networks

Machine Learning
Basic Thoughts

Get to Know what’s going on - Example

Let's assume you are a hardware developer and your board is coming back from prototype manufacturing and it does not work ...
NDI Components

Idea: Analytics Engine

• Universal analysis tool
• Works on every mobile network data
  - Traffica (Nokia Network Data Collector)
  - Service KPIs
  - Log/Trace files
• View concept for specific aspects:
  - Time
  - Radio Cell
  - Parameter correlation
  - Statistics
  - Entity relations
• Applies selectable algorithms: prediction, clustering, regression, (un-)supervised learning, training mode
• Highly interactive
Network Use Cases

KPI Prediction & Time Slot Classification

- Time slot classification on history
  - Normal behavior
  - Outliers
- Long term trend analysis
- KPI radar & prediction
Network Use Cases

Dropped Packet Connections per Radio Cell

- Important SLA criteria
- History Needs to be continuously monitored
- Prediction turns monitoring to preventive activity
- Correlate network problems with 3D buildings
Network Use Cases
Parameter Correlation

• Show as many dimensions as possible
• Show relations between data
KPI Time Series
What does it tell you?

- Drop rates
- Alarm counters
- Performance values

Time
- Resolution: 15min, 1 hour, 1 Day
- N months history

The Problem: A huge amount of simple information

xxx KPIs
x.xxx Hours, 15m intervals
xx.xxx Radio Cells, nodes

KPI - Time

0 .. 24h

0 .. 24h

0
The Problem

... more information does not always help

There must be something more intelligent
KPI Time Series

Operation Challenges

What KPI is operating in (un-)normal state?

Critical situations: When, Where?

What is the reason for that?

What will happen tomorrow?
KPI Time Series

Operation Challenges - Answers Today: Thresholds

Weak Concept ...
• Thresholds depend on time, radio cell ...
• Hard to find the right boundary between good/bad
• Does not reflect seasonality
• Generate a lot of false positives
KPI Time Series

Operation Challenges - Hidden information

Solutions today:
• Experienced ‘eyes’
• But
  • Too much KPIs ... hundreds
  • Too much cells ... thousands
  • Not enough experienced ‘eyes’ ...
  • Error prone

An intelligent system should tell me:
• Where to look at ...
• Where are reasonable anomalies ...
• Where are dependencies to other KPIs, entities
• What are possible reasons for this behavior

Increasing long term trend
(… possible problem in n days)

Possibly a sleeping cell

But this is not a sleeping cell

Anomaly in early hours
KPI Time Series

Some Theory ... Splitting up a Time Series into ...

Split system signal into
- Season component (day, week, month)
- Trend
- Noise

Set upp./low. threshold
- Season component
- Trend

Corridor of expectations

Decreasing Accuracy with increasing forecast window

Accuracy
KPI Time Series

Corridor of ‘expected behavior’

Model violation: Anomaly in early hours

Model violation: Possibly a sleeping cell

Behavior as expected

Trend violation

Corridor of expected behavior
Multi KPI Problems

Finding Anomalies between KPIs

- In most cases the ratio of dropped calls / registered users is in a similar range.
- By using clustering, outliers can be detected.
- Nevertheless, the single KPIs might be in the trusted/expected corridor.
Why Prediction is so Important
Some Facts ...

Business Facts

• Radio cell outage impacts a large amount of customers
• Even if you rapidly realize a problem …. it is always too late ...
• High acquisition costs for new customers

Idea: Change from 'reactive‘ mode to 'preventive‘ mode

• Make an estimation which cells are in trouble next time
• Trigger preventive maintenance actions
KPI Modelling

Now machine learning comes into play

Sizing & computation efforts increase ...
- 1 model per cell & KPI
- x.xxx.xxx number of models

Algorithms

Trend
- Linear Regression

Seasonality
- k Nearest Neighbor or Decision Tree or (non)linear Regression
- Logistic Regression
Problem solving
A Simple Thing

- Find problem
  - Normal operation?
  - Hundred of KPIs
  - Thousands of cells
  - Show deviations

Detect ✓

Explain ?
- Problem location
- Involved systems
- References to
  - Customer
  - Device type

Solve
- Change parameters
- Fix problem
- Invest
Why something happens ...

Root Cause Analysis

Having a KPI is not sufficient ...

- Drop call rate in cell A is too high
- Huge amount of handover failure during busy hours

Reasoning is important ...

- Why does this happen?
- What is driving this problem?
- Is it a periodic or a one time problem?
- What do I need to do, how can I change it?
Root Cause Analysis
Information Complexity – Just a normal case: Handover failures

Questions:
• Errors specifically related to time, BSC, cell, end device
• Any typical errors (DX_CAUSES)

<table>
<thead>
<tr>
<th>Timeslot60</th>
<th>Inter2gTransRealFail</th>
<th>BSC_ID</th>
<th>CELL</th>
<th>DX_CAUSE</th>
<th>LAC</th>
<th>PRB</th>
<th>MM_CAUSE</th>
<th>TAC</th>
<th>HOUR</th>
<th>M15</th>
<th>MG5</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>2014-04-16T14:00:00</td>
<td>1</td>
<td>343</td>
<td>45831</td>
<td>2835</td>
<td>34301</td>
<td>306</td>
<td></td>
<td>01371900</td>
<td>14</td>
<td>0</td>
<td>2</td>
<td>14</td>
</tr>
<tr>
<td>2014-04-16T15:00:00</td>
<td>0</td>
<td>341</td>
<td>43993</td>
<td>778</td>
<td>34100</td>
<td>306</td>
<td>17</td>
<td>35477004</td>
<td>15</td>
<td>3</td>
<td>11</td>
<td>59</td>
</tr>
<tr>
<td>2014-04-16T15:00:00</td>
<td>1</td>
<td>351</td>
<td>56983</td>
<td>2835</td>
<td>35100</td>
<td>306</td>
<td></td>
<td>01264600</td>
<td>15</td>
<td>3</td>
<td>11</td>
<td>59</td>
</tr>
<tr>
<td>2014-04-16T15:00:00</td>
<td>0</td>
<td>342</td>
<td>48171</td>
<td>778</td>
<td>34202</td>
<td>306</td>
<td>17</td>
<td>35750805</td>
<td>15</td>
<td>3</td>
<td>11</td>
<td>59</td>
</tr>
<tr>
<td>2014-04-16T15:00:00</td>
<td>0</td>
<td>343</td>
<td>47493</td>
<td>3328</td>
<td>34300</td>
<td>306</td>
<td>17</td>
<td>35461002</td>
<td>15</td>
<td>3</td>
<td>11</td>
<td>59</td>
</tr>
<tr>
<td>2014-04-16T15:00:00</td>
<td>0</td>
<td>342</td>
<td>40702</td>
<td>778</td>
<td>34201</td>
<td>306</td>
<td>17</td>
<td>01362900</td>
<td>15</td>
<td>3</td>
<td>11</td>
<td>58</td>
</tr>
<tr>
<td>2014-04-16T15:00:00</td>
<td>0</td>
<td>342</td>
<td>42221</td>
<td>2842</td>
<td>34200</td>
<td>306</td>
<td></td>
<td>35491105</td>
<td>15</td>
<td>3</td>
<td>11</td>
<td>58</td>
</tr>
<tr>
<td>2014-04-16T15:00:00</td>
<td>0</td>
<td>343</td>
<td>42593</td>
<td>3336</td>
<td>34302</td>
<td>306</td>
<td></td>
<td>35635505</td>
<td>15</td>
<td>3</td>
<td>11</td>
<td>58</td>
</tr>
<tr>
<td>2014-04-16T15:00:00</td>
<td>0</td>
<td>342</td>
<td>40611</td>
<td>778</td>
<td>34201</td>
<td>306</td>
<td>17</td>
<td>35487901</td>
<td>15</td>
<td>3</td>
<td>11</td>
<td>58</td>
</tr>
<tr>
<td>2014-04-16T15:00:00</td>
<td>0</td>
<td>343</td>
<td>45571</td>
<td>778</td>
<td>34300</td>
<td>306</td>
<td>17</td>
<td>35957605</td>
<td>15</td>
<td>3</td>
<td>11</td>
<td>58</td>
</tr>
<tr>
<td>2014-04-16T15:00:00</td>
<td>0</td>
<td>342</td>
<td>40551</td>
<td>778</td>
<td>34202</td>
<td>306</td>
<td>17</td>
<td>35976050</td>
<td>15</td>
<td>3</td>
<td>11</td>
<td>58</td>
</tr>
</tbody>
</table>

3.624 messages
Root Cause Analysis
Let’s solve it graphically

- Must be an interactive drill down
- Reduces xx.xxx messages to a single problem
- Shows dependencies, anomalies
- Provides possible reasons

Failure rate increases over the day
Peek weekday
Specific cells are more involved
Specific cells are more involved

Specific BSC Problem?
Root Cause Analysis

Solutions: Let’s use an algorithm for that

- **Strong time dependency**
- **Involvement of BSCs**

Show dependencies for error type

**CLEAR/A ONHOOK DURING SET-UP PHASE**

<table>
<thead>
<tr>
<th>Feature</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>30.094</td>
</tr>
<tr>
<td>HOUR</td>
<td>12.699</td>
</tr>
<tr>
<td>M05</td>
<td>5.611</td>
</tr>
<tr>
<td>M15</td>
<td>1.537</td>
</tr>
<tr>
<td>PRB=306</td>
<td>1</td>
</tr>
<tr>
<td>MM_CAUSE=17</td>
<td>0.999</td>
</tr>
<tr>
<td>BSC_ID=341</td>
<td>0.283</td>
</tr>
<tr>
<td>BSC_ID=342</td>
<td>0.247</td>
</tr>
<tr>
<td>BSC_ID=343</td>
<td>0.228</td>
</tr>
<tr>
<td>LAC=34100</td>
<td>0.114</td>
</tr>
</tbody>
</table>

Lines: 41
## Algorithmic View

<table>
<thead>
<tr>
<th>Report types</th>
<th>Sub types</th>
<th>Algorithm</th>
<th>Use case</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model deviation (single field)</td>
<td>Difference between model and real value above limit (x%, configurable)</td>
<td>Nearest Neighbor (KNN)</td>
<td>Unexpected value ‘sleeping cell’</td>
</tr>
<tr>
<td></td>
<td>Difference between model &amp; real value but also value near zero</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trend violation (single field)</td>
<td>Raised if trend slope above/below limit</td>
<td>Linear Regression</td>
<td>Long term trend analysis</td>
</tr>
<tr>
<td>Classification error (multiple field)</td>
<td></td>
<td>DBSCAN</td>
<td>Detecting irregular pattern KPI (combinations not seen before)</td>
</tr>
<tr>
<td>Root cause analysis (multiple field)</td>
<td>Gaussian Naive Bayes</td>
<td>Gaussian Naive Bayes</td>
<td>Find driving factor for specific dx.causes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Show relations to other attributes (cell, customer, …)</td>
</tr>
</tbody>
</table>
Summary
From Simple KPIs to a guided problem solving workflow...

- Trained Model For expected KPI behavior
- Anomaly Detection
- Problem categorization
- Problem evaluation
- Root Cause Analysis

KPI x1 for Cell y1
KPI x2 for Cell y2
KPI x3 for Cell y3
KPI x4 for Cell y3
How to do Time Series Prediction

Prediction of Time Series (ARMA) – Step 1

### Raw Data

<table>
<thead>
<tr>
<th>Time</th>
<th>KPI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time t</td>
<td>KPI value</td>
</tr>
<tr>
<td>Time t-1</td>
<td>KPI value</td>
</tr>
<tr>
<td>Time t-2</td>
<td>KPI value</td>
</tr>
<tr>
<td>Time t-3</td>
<td>KPI value</td>
</tr>
<tr>
<td>Time t-4</td>
<td>KPI value</td>
</tr>
</tbody>
</table>

### Step 1: Computing Moving Averages

<table>
<thead>
<tr>
<th>Time</th>
<th>KPI</th>
<th>Mov AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time t</td>
<td>KPI value</td>
<td>MA KPI</td>
</tr>
<tr>
<td>Time t-1</td>
<td>KPI value</td>
<td>MA KPI</td>
</tr>
<tr>
<td>Time t-2</td>
<td>KPI value</td>
<td>MA KPI</td>
</tr>
<tr>
<td>Time t-3</td>
<td>KPI value</td>
<td>MA KPI</td>
</tr>
<tr>
<td>Time t-4</td>
<td>KPI value</td>
<td>MA KPI</td>
</tr>
</tbody>
</table>

- **Auto**
- **Regression**
- **Moving**
- **Average**

Moving average window
How to do Time Series Prediction
Prediction of Time Series (ARMA) – Step 2

Step 1: Computing Moving Averages

<table>
<thead>
<tr>
<th>Time</th>
<th>KPI</th>
<th>Mov AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time t</td>
<td>KPI value</td>
<td>MA KPI</td>
</tr>
<tr>
<td>Time t-1</td>
<td>KPI value</td>
<td>MA KPI</td>
</tr>
<tr>
<td>Time t-2</td>
<td>KPI value</td>
<td>MA KPI</td>
</tr>
<tr>
<td>Time t-3</td>
<td>KPI value</td>
<td>MA KPI</td>
</tr>
<tr>
<td>Time t-4</td>
<td>KPI value</td>
<td>MA KPI</td>
</tr>
</tbody>
</table>

Step 2: Generating lag window

<table>
<thead>
<tr>
<th>Time</th>
<th>KPI</th>
<th>KPI t-1</th>
<th>KPI t-2</th>
<th>KPI t-3</th>
<th>Mov AVG</th>
<th>M AVG t-1</th>
<th>M AVG t-2</th>
<th>M AVG t-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time t</td>
<td>KPI value</td>
<td>KPI value</td>
<td>KPI value</td>
<td>KPI value</td>
<td>MA KPI</td>
<td>MA KPI</td>
<td>MA KPI</td>
<td>MA KPI</td>
</tr>
<tr>
<td>Time t-1</td>
<td>KPI value</td>
<td>KPI value</td>
<td>KPI value</td>
<td>KPI value</td>
<td>MA KPI</td>
<td>MA KPI</td>
<td>MA KPI</td>
<td>MA KPI</td>
</tr>
<tr>
<td>Time t-2</td>
<td>KPI value</td>
<td>KPI value</td>
<td>KPI value</td>
<td>KPI value</td>
<td>MA KPI</td>
<td>MA KPI</td>
<td>MA KPI</td>
<td>MA KPI</td>
</tr>
<tr>
<td>Time t-3</td>
<td>KPI value</td>
<td>KPI value</td>
<td>KPI value</td>
<td>KPI value</td>
<td>MA KPI</td>
<td>MA KPI</td>
<td>MA KPI</td>
<td>MA KPI</td>
</tr>
<tr>
<td>Time t-4</td>
<td>KPI value</td>
<td>KPI value</td>
<td>KPI value</td>
<td>KPI value</td>
<td>MA KPI</td>
<td>MA KPI</td>
<td>MA KPI</td>
<td>MA KPI</td>
</tr>
</tbody>
</table>
## Results

### Confusion Matrix

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
<th>Predicted as Positive</th>
<th>Predicted as Negative</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neural Networks</td>
<td>93,944</td>
<td>Real Positive</td>
<td>True Positives:</td>
<td>False Negatives:</td>
<td>0.939</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Real Negative</td>
<td>False Positives:</td>
<td>True Negatives:</td>
<td>?</td>
</tr>
<tr>
<td>Non Linear Regression</td>
<td>98,681</td>
<td>Real Positive</td>
<td>True Positives: 12099</td>
<td>False Negatives: 78</td>
<td>0.992</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Real Negative</td>
<td>False Positives: 93</td>
<td>True Negatives: 692</td>
<td>0.899</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>98.388</td>
<td>Real Positive</td>
<td>True Positives: 12155</td>
<td>False Negatives: 62</td>
<td>0.988</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Real Negative</td>
<td>False Positives: 147</td>
<td>True Negatives: 638</td>
<td>0.911</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>96.607%</td>
<td>Real Positive</td>
<td>True Positives: 11836</td>
<td>False Negatives: 143</td>
<td>0.976</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Real Negative</td>
<td>False Positives: 287</td>
<td>True Negatives: 409</td>
<td>0.741</td>
</tr>
<tr>
<td>Ensemble Trees</td>
<td>97.338</td>
<td>Real Positive</td>
<td>True Positives: 12087</td>
<td>False Negatives: 90</td>
<td>0.979</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Real Negative</td>
<td>False Positives: 255</td>
<td>True Negatives: 530</td>
<td>0.855</td>
</tr>
</tbody>
</table>

### KPI:
- **PS_ACC_FA_RT_GCR**
- Pos./Neg. boundary: 6%
- Number of cells: 233 (out of top BTS with highest voice traffic)
- MA Window: 6 hours
- Lag time window: 48 hours
- Prediction window: 1 hour
- Ind. test data: 12177 Positives / 785 Negatives (=KPI violation)

### Results

**Confusion Matrix**:
- **True Negatives**: Correctly predicted KPI violations
- **False Negatives**: Dummy KPI violations incorrectly show as violation
- **False Positives**: KPI violations not found

**Pos./Neg. boundary**: 6%

**Number of cells**: 233 (out of top BTS with highest voice traffic)

**MA Window**: 6 hours

**Lag time window**: 48 hours

**Prediction window**: 1 hour

**Ind. test data**: 12177 Positives / 785 Negatives (=KPI violation)
Network Data Intelligence
Research Areas

**Applications**
- Customer experience management
- Churn prediction
- Customer segmentation
- Geo Analytics
- Social Analytics

**Use Case**
- Customer experience management
- Churn prediction
- Customer segmentation
- Geo Analytics
- Social Analytics

**Methods**
- Data mining
- Machine Learning
- Statistical Analysis
- Predictive Analysis
- Tools

**Storage**
- Data warehouses
- NoSQL Databases
- Column stores
- In memory DBs

**Processing**
- Hadoop engines
- Real time analytics

**Representation**
- Dashboards
- Data visualization
- Rich clients
- Collaboration platforms

**Use Cases**

**Technology Enabler / Platform component**

© Nokia Networks 2015
Generalized Big Data Analytics Stack

What do you need?

- Use Cases
- Visualization, Charting, Drill Down Views
- Analytics Algorithms (K-Means, KNN...)
- Data Storage (Relational, NoSQL)
- Aggregation, Filtering, Distributed Computing
- Import, Formatting, Type Conversion
- Data Sources: net elements, protocols
Generalized Big Data Analytics Stack

How does it correspond to the efforts?

Use Cases
Visualization, Charting, Drill Down Views
Analytics Algorithms (K-Means, KNN...)
Data Storage (Relational, NoSQL)
Aggregation, Filtering, Distributed Computing
Import, Formatting, Type Conversion
Data Sources: net elements, protocols

Development Effort
Generalized Big Data Analytics Stack

Software Choices

Analytic Tools
- RapidMiner
- Knime
- SPSS
- Tableau, QlikView
- Mahout
- Hive
- Parallel processing
  - Hadoop
  - Spark
  - Storm
- Pig
- Hive
- Pandas
- SCI-Kit
- R
- Python
- Sci-Kit
- Pandas
- Python
- R

Language Stacks
- Big Data DBs
  - Oracle
  - Teradata
  - NoSQLs
  - Vertica

Use Cases
- Visualization, Charting, Drill Down Views
- Analytics Algorithms (K-Means, KNN...)
- Data Storage (Relational, NoSQL)
- Aggregation, Filtering, Distributed Computing
- Import, Formatting, Type Conversion
- Data Sources: net elements, protocols

Requires good analyst background

Needs to be complemented by powerful DB

Problems with Big Data

Not very popular

High entry barrier

Crowded place

Needs to be complemented by powerful DB

Problems with Big Data

Not very popular

High entry barrier

Crowded place

Requires good analyst background

Needs to be complemented by powerful DB

Problems with Big Data

Not very popular

High entry barrier

Crowded place
Only a few corresponding alarms with KPIs above limit

Very weak alarm status
Generalized Data Analytics Stack

Final Choices

Use Cases
Visualization, Charting, Drill Down Views
Analytics Algorithms (K-Means, KNN...)
Data Storage (Relational, NoSQL)
Aggregation, Filtering, Distributed Computing
Import, Formatting, Type Conversion
Data Sources: net elements, protocols

Analytic Tools
RapidMiner
Knime
SPSS

Tableau, QlikView
Mahout
Hive
Parallel processing
Hadoop, Spark, Storm
Pig

Language Stacks
SCI-Kit
(Phoenix)
R
Python
Pandas
(Phoenix)

Big Data DBs
Oracle
Teradata
NoSQLs
Vertica
Generalized Data Analytics Stack
NDI Software Architecture

Data Sources:
- Network Element Data
- DPI Data
- Service KPIs
- Enrichments
  - OpenGML, TAC, Locations

NDI – Client
NDI – Server
NDI – Distributed Real Time Importer

Use Cases
- Visualization, Charting, Drill Down
- Analytics Algorithms (K-Means, KNN...)
- Data Storage (Relational, NoSQL)
- Aggregation, Filtering, Distr.Comp.
- Import, Formatting, Type Conversion
- Data Sources: net elements, protocols

SQL DBs
NoSQL DBs
Network Data Intelligence Demonstrator
NDI - Detailed Architecture

Browser Client
- OpenLayers
- HighCharts
- D3
- Angular
- OceanTouch
- JavaScript
- HTML
- CSS

Server
- Django
- Python
- SCI Kit
- Pandas
- RapidMiner Server

Database Layer
- SQL
- MongoDB
- MySQL
- Others

Development Environment
- RapidMiner
- Knime
- Tableau
- QlikView

Real Time Streaming
- Pandas
- SCI Kit
- Python

RAW Data
- Artificial Data Generator

HTTP REST

Standard Programming
Data Analytics & Aggregation
Rich Client & Charting
Tool
Network Data Intelligence

Reasons to use Python

- **Rapid Prototyping**
  - Interpreted, very dense coding, eclipse supp. (PyDev)
  - Zero turnaround, no SW production

- **High Functionality**
  - Data analytics, Statistical computing, Text mining
  - Modern ‘R’

- **Next generation programming**
  - Object oriented, Functional programming
  - Closures, Duck typing, memory management, Lambdas

- **Huge library & community**
  - Uses most C-libs on Linux - therefore very fast
  - Communication stacks, encryption, ...

- **Support for server programming**
  - Django (HTTP based)
  - Twisted (event based)

---

Google: “... we use Python, where we can, C/C++, where we must ...”
## Experiences / Details

### Python - Possible Counter Arguments

<table>
<thead>
<tr>
<th>Topic</th>
<th>Details</th>
</tr>
</thead>
</table>
| **Speed**                                  | • Think big and optimize your algorithms  
• Use full potential of your database                                                                                               |
| **No support for parallel programming (GIL)** | • Use secure message passing                                                                                                         |
| **Python 2/3 incompatibility**             | • That’s more or less done  
• Most packages have Python 3 support today                                                                                     |
| **Unstable – nothing for product development** | • That’s just wrong  
• Our code just worked from the beginning, no GC pain, no bugs, ...                                                                 |

Our programmers love it
Important NDI Components
Pandas

- R-extension for Python
- ‘in-memory’ SQL
- Fast – native C-arrays
- Data types:
  - Series, DataFrame, Panel, 4D
- Vector operations
- IO operations (CSV, DBs, …)
- Descriptive statistics
- Group by, sort, indexing
- Merge, join, concatenate
- Reshape, pivoting
- Time series analysis

SCI-Kit

- (Un)-Supervised learning
  - Decision trees, ...
- Classification
  - SVM, nearest neighbors, random forest, ...
- Clustering
  - k-Means, spectral clustering, mean-shift, ...
- Regression
  - SVR, ridge regression, Lasso, ...
- Data pre-processing
  - Normalization
Network Data Intelligence
Importer Data Model

- Describes transformation of raw data to a n-dimensional OLAP cube
- Dimensions
  - 1st aggregation dimension
    - CGI, IMSI, URL, IMEI, Service ...
  - 2nd aggregation dimension
    - Time
  - 3rd aggregation dimension on demand
    - IMSI / CGI (e.g. for mobility)
- Aggregation levels
  - Time: 15min, hour, day, week, month
  - CGI: MCC, MNC, LAC, Class ID
  - IMEI: TAC, SNR
  - IMSI: MCC, MNC, MSIN
Data Importer Architecture

Scan
- Tree Walker
- ZIP Reader
- File Reader
- DB Reader
- URLReader
- Socket Reader

Compute
- Pandas
  - Data Type Conversion
  - Filtering / Cleaning
  - Aggregate
  - Join / Pivoting
- SCI-Kit Learn
  - (un)supervised Learning
  - Model Application
  - Data Transformation

Write
- MySQL Writer
NDI Parallel Real Time Engine

- DPI
- Net Element
- Net Element
- KPIs
- Enrichments

Message Broker

- N * Real Time Engine workers
- N * Real Time Engine workers
- N * Real Time Engine workers
- N * Real Time Engine workers
- N * Real Time Engine workers
- N * Real Time Engine workers

NDI Server

- NDI Rich Client

Speed: Up to 150,000 msg/sec on 4 server system
Important Importer Components

Celery

• Distributed task messaging
• Tight python integration
  - @task
  - Django integration
  - Django management support module
• Based on several message brokers
  - RabbitMQ
  - Redis
  - ...others
• Inter / intra node operation

RabbitMQ

• Message Broker
• Communication patterns
  - 1:1, 1:n, n:1
  - Publish / subscribe
• Multiple queue support
• Complex message routing
• AQMP standard
• Multi language support
• Written in Erlang
Rich Client Development

Basic Thoughts about Browsers ...

Virtual machines

- Container for complex applications
- Standardized environment (better than any alternative)

... with multiple DSL support

- HTML5: DOM structure
- CSS3: look & feel
- JavaScript: behavior

... comes with development environment

- Extremely fast and well tested
- Available everywhere: PCs, Tabs, Mobiles
- Powerful choices: Chrome, Firefox (not IE)

And they are for free
Rich Client Development

Web application types

**Client rendering**
- JavaScript application loaded via HTTP
- Requires client framework
  - Angular, Ember
- Complex JavaScript application
- Triggers locally handled
- Fast local interaction
- Data access via JSON
- Lots of JavaScript
- Behaves like an application

**Server rendering**
- Fully rendered web page downloaded
- Requires server framework
  - Java: GWT, JSF, Portlets ...
- Pure server programming
- All interactive events across server
- No local interaction
- Data embedded in DOM
- Mostly server programming
- Less interactive
Programming with JavaScript

A reasonable language for application development

No. 8 in Tiobe index

Standardized (ECMA 5..6)

Functional / object oriented, Closures

Dynamic Typing, Interpreted & broadly available

Huge amount of upcoming frameworks: Charting, Angular ...
Rich Client Development

Client Components/Libraries

Openlayers
- All kind of maps (OSM, Google ...)
- Graphics layers

HighCharts
- Complex charts
- Highly interactive (zoomin/shifting, event handling, ...)

D3
- Complex graphics (Force directed graphs, trees ---)
- Low level graphics and event handling
Rich Client Development

The Need for Angular

- Angular capabilities
  - MVC architecture
  - 2 way binding
    - JavaScript <-> HTML/DOM
  - Expressions
  - Directives
    - ng-show, ng-repeat
  - Routing
    - Introduces sub commands to URL

- ‘Tomcat / JBOSS’ for rich client development
- Great simplification for development
- Reduces amount of necessary JavaScript code
Big Data Analytics in Telecommunication

Wrap Up
Thank You!
norbert.kraft@nokia.com