An Adaptive Scalable Analytics Platform

ASAP: **A**daptive, highly **S**calable **A**nalytics **P**latform)

EC project: FP7-ICT-2013-11 ASAP 619706

Description of the initial dataset

Deliverable no.: D9.1
Date: 31 March 2014

ASAP is funded by the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement No: 619706

Table of Contents

Deliverable title:	Description of the initial dataset
Filename:	asap_d9-1_31mar2014-5.docx
Author(s)	Roberto Bertoldi, Fabio Pandini, Rita Spada
Date	31 March 2014

Start of the project:	01 March 2014
Duration:	3 years (36 motnhs)
Project coordinator organisation:	FORTH

Deliverable title:	Description of the initial dataset
Deliverable no.:	9.1

Due date of deliverable:	M1
Actual submission date:	31 March 2014

Dissemination Level

X	PU	Public
	PP	Restricted to other programme participants (including the Commission Services)
	RE	Restricted to a group specified by the consortium (including the Commission Services)
	CO	Confidential, only for members of the consortium (including the Commission Services)

Deliverable status version control

Version	Date	Author
1.0	31 March 2014	R. Bertoldi, F. Pandini, R. Spada

Abstract

The present deliverable describes the initial dataset provided by WIND Telecomunicazioni SpA (WIND) at the beginning of the project. We also briefly describe the structured data used for customers information storage(CRM) and the semi-structured data used for services charging information storage (CDR). In addition a brief description of the data base infrastructure used for data storage is given along with some data structure examples.

Keywords

Customer Relationship Management (CRM), Charging Data Records (CDR), structured data, semi-structured data, unstructured data, unstructured data, social networks, Facebook

1 WIND Data Centre infrastructure

1.1 Overview

The present deliverable describes the initial dataset provided by WIND at the beginning of the project. We also briefly describe the structured data used for customers information storage (CRM) and the semi-structured data used for services charging information storage (CDR). In addition a brief description of the data base infrastructure used for data storage is given along with some data structure examples.

1.2 WIND Information Technology infrastructure

In the **Tables** below a summary of WIND Data Centre infrastructure deployed capacity is given.

DATA	
Storage boxes	**126**
Disk Capacity (PB)	**6.20**
Transactions/s	**120,000**
MB/s	**1000**
Backup (TB/day)	**223**

APPLICATIONS	
Main applications	**185**
System availability	**99,98 %** (YTD 2013 actual)

TOOLS	
Personal Computers	**6500 +**
Thin Clients	**1300**
Hardware models	**112**

SYSTEMS	
Servers	**2500 +**
Operating Systems	**30 +**
Databases	**1000 +**

CONNECTIVITY	
Appliances (routers, switches, etc.)	**1200 +**
Firewalls	**38**
LAN active ports	**17,000 +**
Installed cables	**470 km +**
Installed fibres	**750 km +**

The available Data Centres and associated Server Farms are deployed in three sites across Italy as illustrated in the **Figure** and **Table** below:

Geographic Distribution of the Data Centres and Server Farm

Data Center	**Ivrea "XXV Aprile"**	**Molfetta**	**Milan "Ortles" (Server Farm)**
Size (m^2)	**2000**	**2400**	**550**
Power capacity max (MW)	**5**	**3.25**	**1.25**
Cooling capacity max (BTU)	**19,885,000**	**3,488,220**	**5,118,211**
Production	**X**	**X**	**X**
Development		**X**	**X**
Test		**X**	**X**

1.3 WIND proposed architecture for ASAP

The Figure below shows WIND's proposed architecture for ASAP. The Figure below shows WIND's proposed architecture for ASAP.

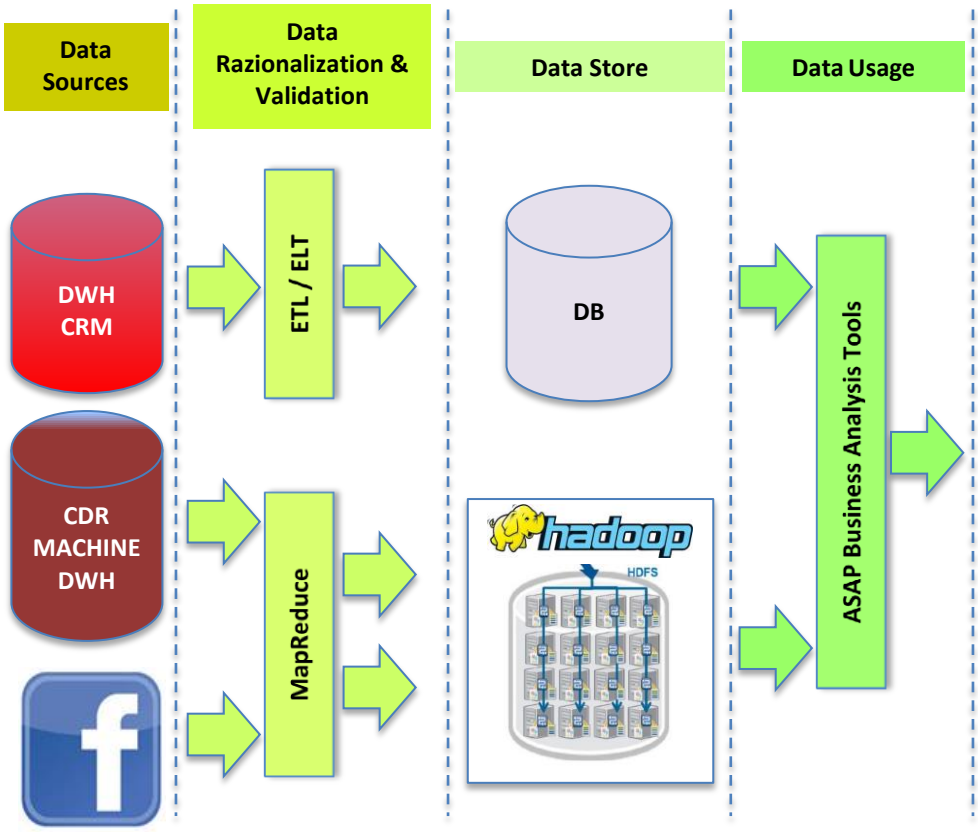

Figure 1: WIND proposed architecture

2 WIND Dataset

The foreseen dataset to be used in the ASAP project is briefly listed and described in the following paragraphs.

We will consider the following types of data as:

- Structured: **Customer Relationship Management (CRM)** data typically related to the clients personal information.
- Semi-Structured: **Charging Data Records (CDR)** related to Voice, SMS and Mobile Traffic data used for billing typically.
- Unstructured: "mined" from **Facebook** posts present on defined/specific **"Facebook Fan Pages"**

2.1 Customer Relationship Management (CRM) data

The CRM data source holds the customers personal data. The average size of this dataset is of about 1.5 million customers. The data will be fully anonymized prior to being used in order to preserve the privacy of the customers as required by the current national privacy protection laws. For the ASAP project this dataset can be used to select specific customers who reside in an area of interest for statistical and/or measurement purposes. The customers/users are in principle identified by a relationship key given by their Mobile Subscriber Identity Number (MSISDN).

2.2 Charging Data Records (CDR) data

The CDR data that hold the customers billing information and/or traffic data. For the analysis purposes of the ASAP project CDR data can be collected historically and this dataset has an average of 3 billion records (for 3 months) and about 1 billion records for the current month. Again, as in the case of CRM data (see above) these data will be fully anonymized prior to being used in order to preserve the privacy of the customers as required by the current (national) privacy protection laws.

The typical data structure of CDR data is shown in the **Table** below:

FIELD NAME	TYPE	DESCRIPTION
SESSION_ID	NUMBER	Incremental sequence number
ID_CALLER	NUMBER	Caller ID identifier. Encrypted data.
ID_CELL_START	VARCHAR2(14)	ID of the caller cell at call start
TS_START	DATE	Call start timestamp
ID_CELL_END	VARCHAR2(14)	ID of the caller cell at call end
DURATION	NUMBER	Call duration in seconds

As an example the raw contents of a CDR record is shown below:

…
11475150708|2012-07-10|074850|370|A81E704D|VIA CARLO BILOTTI , 43|87100|A81E757F|VIA PANEBIANCO , -|87100
11475150708|2012-07-10|171109|970|A81E757F|VIA PANEBIANCO , -|87100|A81E757F|VIA PANEBIANCO , -|87100
11475150708|2012-07-10|185803|480|A81E757F|VIA PANEBIANCO , -|87100|A81E757F|VIA PANEBIANCO , -|87100
11475150708|2012-07-10|194952|30|A81ED4D9|VIA CARLO BILOTTI , 43|87100|A81E704D|VIA CARLO BILOTTI , 43|87100
11475150708|2012-07-10|223345|2470|A81E757F|VIA PANEBIANCO , -|87100|A81ED4D9|VIA CARLO BILOTTI , 43|87100
11475150708|2012-07-11|104124|1340|A81E704D|VIA CARLO BILOTTI , 43|87100|A81E704D|VIA CARLO BILOTTI , 43|87100
11475150708|2012-07-11|114343|1110|A81ED4D9|VIA CARLO BILOTTI , 43|87100|A81ED4D9|VIA CARLO BILOTTI , 43|87100
11475150708|2012-07-11|161909|1470|A81E757F|VIA PANEBIANCO , -|87100|A81E757F|VIA PANEBIANCO , -|87100
11475150708|2012-07-11|165536|490|A81E704D|VIA CARLO BILOTTI , 43|87100|A81E704D|VIA CARLO BILOTTI , 43|87100
11475150708|2012-07-11|183019|3470|A81E757F|VIA PANEBIANCO , -|87100|A81E757F|VIA PANEBIANCO , -|87100
11475150708|2012-07-11|214756|380|A81E704D|VIA CARLO BILOTTI , 43|87100|A81E757F|VIA PANEBIANCO , -|87100
11475150708|2012-07-11|215149|140|A81E704D|VIA CARLO BILOTTI , 43|87100|A81E704D|VIA CARLO BILOTTI , 43|87100
…

2.2.1 Data selection criteria

Based on the different selection criteria required, specific fields of CDR data will be used according to the statistical analysis desired or for the Social Network Analysis (SNA) purposes as summarized below. Notice that these data will still be related with the original CRM data although in a completely/irreversible anonymous way.

Data: main CDR definition fields
Selection Criteria 1 (for SNA purposes):
They will be related to the CRM of the clients object of analysis
Selection Criteria 2 (for statistical analysis):
They will be related to the CRM of the clients object of analysis
They will have to be generated in the area of interest and in the timeframe of interest
Selection Criteria 3 (for statistical analysis):
They will have to be generated in the area of interest and in the timeframe of interest regardless that they are related or not to the customers subject to analysis
Relationship key: MSISDN

2.3 Social Network (SN) data

For the purpose of social network influence analysis a link between the CDR and CRM data and the “I Like” in Facebook Fan pages will be used and/or incorporated.

The average data generated from the Facebook pages that can be collected Historically (for 3 months) is about 6000 posts while it amounts to about 2000 posts for the current month.

2.3.1 Social Network data analysis

Data: analysis of the “I Like" on the interested “Fan page”
Selection Criteria (for statistical analysis) :
They should not be related in any way to the CRM of the clients
They should not contain any indication of personal data
Relationship key: none