Wiki Operacional PSafe

# Project Objective

We increasingly live in a world where data doesn't fit nicely into the tidy rows and columns of an RDBMS. Mobile, social, and cloud computing have spawned a massive flood of data. According to a variety of estimates, 90 percent of the world's data was created in the last two years, with Gartner pegging 80 percent of all enterprise data as unstructured. What's more,semi-structured [http://wiki.psafe.com/doku.php?id=documentacao:bi:data_dictionary&#semi-structured_data] and unstructured data [http://wiki.psafe.com/doku.php?id=documentacao:bi:data_dictionary&#unstructured_data] are growing at twice the rate of structured [http://wiki.psafe.com/doku.php?id=documentacao:bi:data_dictionary&#structured_data] data.

PSafe has 25MM++ users and recieves 100+ GBs of semi strucuted data every day. The company demands a modern data architecture, to process all this information. This is the reason of this project, prepare PSafe for actual and future challenges.

**The major objective of the project is to replace actual Hadoop cluster with better Data Architecture, what will allow the company achieve business objectives.**

# Solution Description

This project was started in Nov, 2014 and was motivated by current situation issues and new business demands. To do so, we had to migrate old hadoop cluster to a new data architecture, with more efficient ETL process and data storage.

The biggest change was to move Hadoop role, from simple landing zone to a Data Hub [http://wiki.psafe.com/doku.php?id=documentacao:bi:architecture#data_hub_-_future_state], where data is queryable in apropriate Data Layers [http://wiki.psafe.com/doku.php?id=documentacao:bi:architecture#data_layers].

In a very high level, we will use Hadoop for Landing Zone, Staging Area, ODS and Archiving. Data Werehouse will continue on MSSQL and DMs on SSAS.

The BI Architecture Style of this project is Continuous Intelligence: Large volumes of fast changing data, operational process support, collect-monitor-analyse services, complex event processing [http://wiki.psafe.com/doku.php?id=documentacao:bi:data_dictionary&#cep] and event streaming systems.

Please always refer to the PSafe Data Dictionary to complete the understanding of this document.

One of the most fundamental decisions to make when architecting a solution on Hadoop is determining how data will be stored in Hadoop, since it has no standard data storage format.

## Business Drivers

Decision makers can ask themselves the following questions to gauge the need for big data technology:

- Are the current data sets large? Are you limited by your current platform or environment because you can't process the amount of data that you want to process?
- Is the existing warehouse environment a repository of all data that is generated or acquired?
- Do you have much cold or low-touch data that is not being used to analyze and derive business insight?
- Do you want to be able to analyze non-operational data?
- Do you want to use your data for traditional and new types of analytics?
- Are you unable to analyze new sources of data because the data does not fit neatly into schema-defined rows and columns without sacrificing fidelity or the rich aspects of the data?
- Do you need to ingest data as quickly as possible? Does your environment require generation of the schema during run time?
- Are you looking for ways to lower your overall cost for analytics?
- The situations that are described by these questions can be improved by augmenting the existing data warehouse environment with big data technologies.

**We must generate USD$0.30 per month per MAU if we want a profitable cohort (positive LTV). Not an easy task, since we are at about 0.06 today. A lot of work to do here…. Marco, PSafe CEO, nov 17 2014. This must be our mantra at PSafe!!**

## Business Objectives

So there are executive questions to be answered, business financial objective to be reached. Data Team is in charge to start the transformations to make PSafe a data driven company.

First wave of Big Data Solutions mostly cared about 2 of the 3 Big Data V's definition. Companies were worried to manage Volume, Velocity and viraity of the data. Now Psafe joins the second wave, when veracity becames a competitive advantage: How quickly the process can move that data trough all analytics life cycle, understand what the data says and act on response to insights. That's a Data Driven Company.

**Get More data, faster, and act on it.**

**Acquire data faster, analyse data fastar, act on that data faster.**

This new architecture is the base of all this change. The Business data objectives are:

- Real time campaign optimization
- Real time user profile calculation
- Customer experience monitoring
- 360 degrees analysis, complete data integration for Lifetime, Retention, Cost and Revenue.

We strongly belive this solution is the best choice for PSafe challenges.

## PSafe OKRs

1) **Android Market Share (Installs): 51% in Brazil and 25% in the rest of LatAm* by year-end 2015**
   - Key result 1: At least 2 new products / apps launched for android over the next 6 months
   - Key result 2: MAUs, DAUs, and retention metrics for users aligned with company Financial Plan
2) **Android Organic Profile: 40% by year-end 2014; 55% by year-end 2015**
   - Key result 1: Facebook fans / likes and volume of social sharing through the app(s)
   - Key result 2: Become associated with "protection for my digital / connected life" in Brazil
3) **Financial: EBITDA positive by Q3 2015 and cash-flow positive by Q4 2015**
   - Key result 1: Revenue and marketing costs aligned with company Financial Plan
   - Key result 2: Opex, Capex, headcount, cash balance aligned with company Financial Plan
4) **Team: Bring top talent to PSafe to face our challenges and deliver on our OKRs**
   - Key result 1: Open/grow China office to expand dev team and design new apps
   - Key result 2: Grow PSafe Brazil team with top performers in Android and other areas

**Data volume increase is expected, new information requests are coming every day. BI Architecture must support FDE (fast data exploration), BPM (business process monitoring), Real time BI and Self Service BI.**

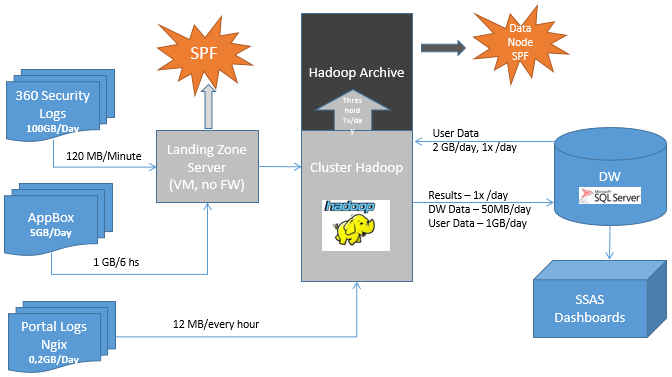
### Investments

This project was implemented with minimun investments, the idea was to reuse hardware, software and human existing resources. Leverage of existing technologies, even user facing and operational facing. The biggest investment was the accisition of a swith juniper qlx3500 10G FC, so we can reach 10GB/s communication speed between data nodes of the cluster.

Existing hardware and software, open source solutions and experienced team are the key ideas to keep costs as low as possible. Please check co-existence section to check HW migration from one cluster to the other.

## Current State

**Data Flow**



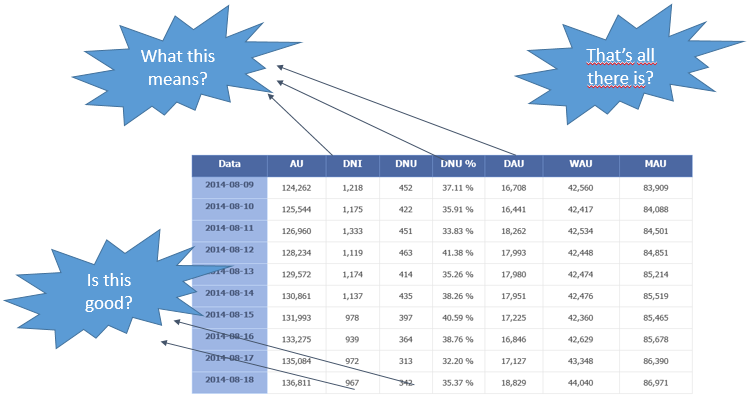

**Data Flow Issues:**

- Raw data "hidden" in raw files, folders, java map reduce jobs required for any access.
- Single Points of Failure.
- Slow data integration.
- Small delivery capacity.
- Confused multi directional data flow.
- Business rules enforced in query time (map reduce time).
- Bad performance and administration.
- Outdated Hadoop Version - all eco system.
- Limited network performance between data nodes within Hadoop Cluster.

- Lack of Audit, Balance and Control.
- ETL processes without exception treatment, all or nothing.
- ETL processes highly connected one each other, dificult maintenance.

Data team created this project to address all this problems and transform the way Psafe process information.

**Lack of Metadata**




| Data | AU | DNI | DNU | DNU % | DAU | WAU | MAU |
|---|---|---|---|---|---|---|---|
| 2014-08-09 | 124,262 | 1,218 | 452 | 37.11 % | 16,708 | 42,560 | 83,909 |
| 2014-08-10 | 125,544 | 1,175 | 422 | 35.91 % | 16,441 | 42,417 | 84,088 |
| 2014-08-11 | 126,960 | 1,333 | 451 | 33.83 % | 18,262 | 42,534 | 84,501 |
| 2014-08-12 | 128,234 | 1,119 | 463 | 41.38 % | 17,993 | 42,448 | 84,851 |
| 2014-08-13 | 129,572 | 1,174 | 414 | 35.26 % | 17,980 | 42,474 | 85,214 |
| 2014-08-14 | 130,861 | 1,137 | 435 | 38.26 % | 17,951 | 42,476 | 85,519 |
| 2014-08-15 | 131,993 | 978 | 397 | 40.59 % | 17,225 | 42,360 | 85,465 |
| 2014-08-16 | 133,275 | 939 | 364 | 38.76 % | 16,846 | 42,629 | 85,678 |
| 2014-08-17 | 135,084 | 972 | 313 | 32.20 % | 17,127 | 43,348 | 86,390 |
| 2014-08-18 | 136,811 | 967 | 342 | 35.37 % | 18,829 | 44,040 | 86,971 |

**SWOT Analysis**

| Strength | Weakness |
|---|---|
| Processing Power | Delivery Time |
| Costs | Actual cluster performance and monitoring |
| Self Service BI | IT Dependency |
| Team expertize | Lack of access to row level data |
| | Naming rules |
| | Metadata |
| **Opportunity** | Threat |
| Agile DW | Lack of time |
| Optimized new cluster | Single Points of Failure (SPF) |
| Monitored new cluster | Lack of business rules definitions |
| C-Level support | Lack of ETL flow and data quality management |
| New Data Architecture | |

## Future State

## Architecture Drivers

Technical objectives and definitions of the new PSafe BI Architecture:

- Architecture Landscape

- Use open source or existing tools: Cloudera Ecossystem and MS SQL Server BI Suite.
- Parallel processing.
- Real Time integration.
- Blend actual BI infrastructure with newer technologies.

- Data Flow (ETL)

- Data flows in only one direction, from Data Sources unitl DM in SSAS.
- Reduce Java usage to Logs parsing only.
- Data intensive operations in the data layer: ELT instead of ETL.
- Cold data moved to cheaper storage.
- Parallel Processing.
- Compress data to reduce footprint. Column storage is an option, as we are doing with Hbase, a type of columnar database.
- Follow Hadoop, HBase and Hive Best Practices.
- Each Data Source must have it's own ETL process, independent from the others. Separete deployments, schedules, scripts, jobs.
- So, ODS does not receive data from DW as we have in current state.
- SQL or HQL interface to query all data in all layers [http://wiki.psafe.com/doku.php?id=documentacao:bi:architecture#data_classification_storage_interface_and_retention].
- Incomplete data does not abort ETL process. It is saved to rejected storage. Please check Data Quality [http://wiki.psafe.com/doku.php?id=documentacao:bi:architecture#data_quality].
- Discovery Zone [http://wiki.psafe.com/doku.php?id=documentacao:bi:architecture#data_discovery_reference_architecture]: Exploration capacity to all data layers.

- Raw Files (Landing Zone)

- Original files before parsing.
- Use the same HDFS area for landing zone and archiving - HAR (Hadoop Archiving).
- Do not use Map Reduce for parsing.
- Real Time integration.
- Archivived on offline storage.
- Schema validation on Read - Semi sctructured Data processing, Data Lake.

- Raw Data (Staging Area)

- Raw files after parsing.
- Staged at HDFS/Hive: partitioned, compressed and indexed.
- Having online access to data in its raw, source form — "full fidelity" data — means it will always be possible to perform new processing and analytics with the data as requirements change.
- Archived queryable at HAR.
- Flexible data model.

- Business Data (ODS)

- Raw data after business rules: uniqueness, timestampt conversion to date and time, etc.
- Master and Reference Data: User Profile, Operational Data.
- Sqoop to load data from ODS to MSSQL. Data is bulk loaded into temp tables, another Staging Area. Then all data is integrated with dimensions in Star Schema.

## Business Rules

Major Business Rules to be applied on raw data.

Explanation:

- raw files → Original files stored in HDFS Landing Zone.
- raw data → raw files parsed and stored in Hdfs/Hive Staging Area.
- business data → **raw data + business rules**, stored in Hbase ODS.
- Bi data → business data integrated with BI Dimensions.

For more details, please check Data Layers [http://wiki.psafe.com/doku.php?id=documentacao:bi:architecture&#data_layers] section.

### Appbox Business Rules

- Impressions: uniqueness per IMEI in 1 hour windows.
- Clicks: uniqueness per IMEI in 1 hour windows.
- Installations: uniqueness per IMEI in all times, "ever uniqueness". Some direct contracts pays per Installation.
- App Open: uniqueness per IMEI in all times, "ever uniqueness". Parteners like Glispa pays per app open.
- Server timestamp, not 360 neither device.
- Revenue is calulated per campaign CPI.
- If Campaign is null or nonexistent, it is a log from old version (1.7 or earlier). We should get the latest active campaign for that app on the APPBOX system. Campaing is kept as null. Data should be moved to Partener LEGADO".
- Campaign ID 0 (zero) means free APP - VERY Rare cases.
- If the app has no campaign, client will send campaign id 2099040713246971. That's a free app too.
- There are Campaigns with CPI zero.
- We need to create all existing KPIs per campaingn country. Phase 2.
- Automatic Campaigns, type 2, have data coming from partner's feed. Including CPI, in Dec 3rd 2014 there is on campaign with CPI = US$ 2100,00. For now, all CPIs above US$10,00 should be devided by 1000. So 2000,00 will be 2,00 .
- Campaign data must be integrated from Appbox Mgt System every 30 minutes.
- eCPC: Effective Cost Per Click. Formula: (Package.Payout * Qtd_Install) / Qtd_Download)
- eCPM: Effective Cost per Mille. Formula: (Revenue / Impressions) + 1000
- CTR: Click trought Rate. The Formula: CTR = clicks/impressions
- Install Rate: Formula = installations / clicks

### Android Business Rules

- All times uniqueness, per IMEI.
- IMEI is encrypted with MD5 algorithm.
- Server timestamp, not 360 neither device.
- Publisher and SubPublisher from MAT data integration.
- Integrate with MAT in this order: 1st) google ID. 2nd) IMEI. 3rd) Android ID.
- Fraud detection: Compare 360 country, IP Country, MAT Country and Campaign Country. At least Campaign and IP must match. There is no Campaign outside LATAM. For now: Brazil, Colômbia, México and Argentina.
- Fraud Detection: Total installations from the same sub net. If bigger than 1000, must be investigated.
- DAU: Daily Active Users (unique). Devices with at least one access in the day.
- DNI: Daily New Installations (unique). Upgrades doen't count.
- DEU: Daily Engaged Users (Unique). Device users that really used the APPBOX in that day.
- Inactive: If any IMEI installed has NOT connected to the base for over 30 CONSECUTIVE DAYS it is considered to be INACTIVE. This is Marco's definition, emailed at 13/nov/2014.

## Co-existence

We will have 1 month period with both old and new cluster in production at the same time. For now, distcp routine (runs every 3 hours). But since 1st Wave, below data source folders must be transfer for both job machines:

- /mobile_operational_data/
- /mobile_appbox/
- /pc_operational_data/
- /security_service_data_mobile/

The servers of the old cluster will be decommissioned as possible, reformatted and added to the new cluster. Old cluster needs to keep on running to support LockBox and existing reports. First wave will move 4 nodes from old to new cluster, one named node and three data nodes.

Migration must be done by DW fact table. Need to use the checklist below.

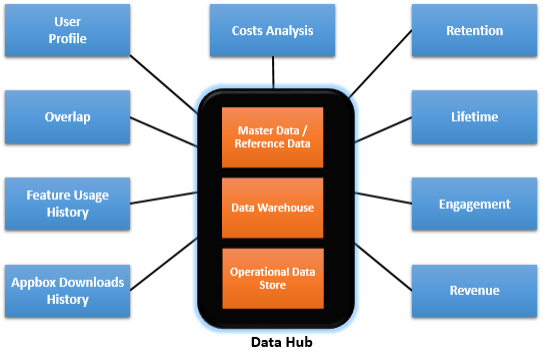
| Fact Table | Migrated |
|---|---|

| Fact Table | Migrated |
|---|---|
| AppBox_FactGInfo | |
| AppBox_FactStatistics | |
| AppBox_NEW_FactPositions | |
| AppBox_NEW_FactStatistics | |
| AppBox_NEW_FactStatistics_Country | |
| AppBox_NEW_FactStatistics_Country_Partner | |
| FactActivePCs | |
| FactAnActivePCs | |
| FactAnFeatures | |
| FactAnInstallations | |
| FactAnRetention | |
| FactAnRetention_Installation | |
| FactAnSubPublishers | |
| FactChannelRetention | |
| FactErrors | |
| FactInstallations | |
| FactReinstallations | |
| FactScans | |
| FactStatistics | |
| FactUninstallations | |
| FactUpgrades | |
| FactWinFeatures | |
| MAT_FactInstallation | |
| Overlap_FactAnOverlap | |
| Overlap_FactBootOP | |
| Portal_FactArticles | |
| Portal_FactClick | |
| Portal_FactPartners | |
| Portal_FactPartnerTraffic | |
| Portal_FactRejectionRate | |
| Portal_FactVisualization | |

## Data Hub

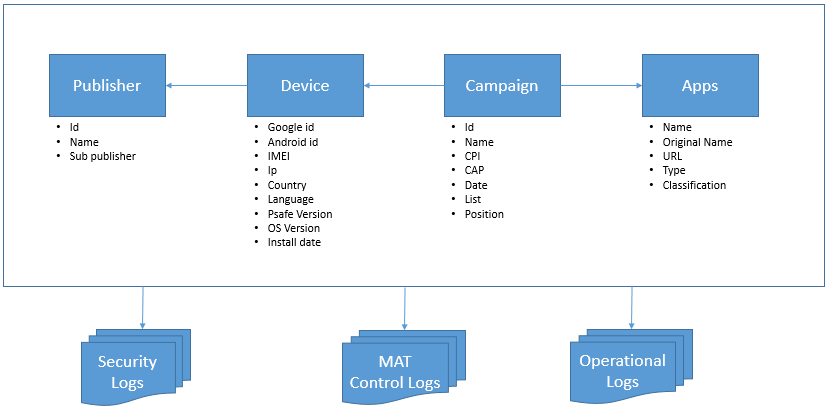
### Enterprise Data Model

The Information Hub provides the data foundation to support the customer interaction and transaction data, that enables a comprehensive single view of the customer throughout the enterprise.

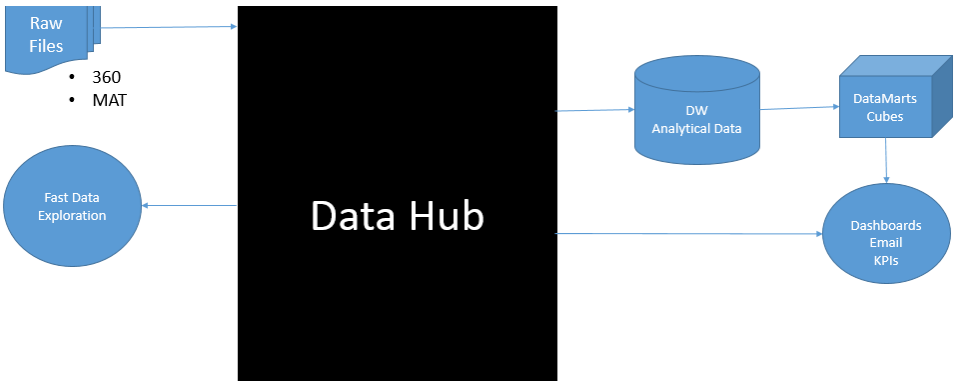



### Business Dimensional Model

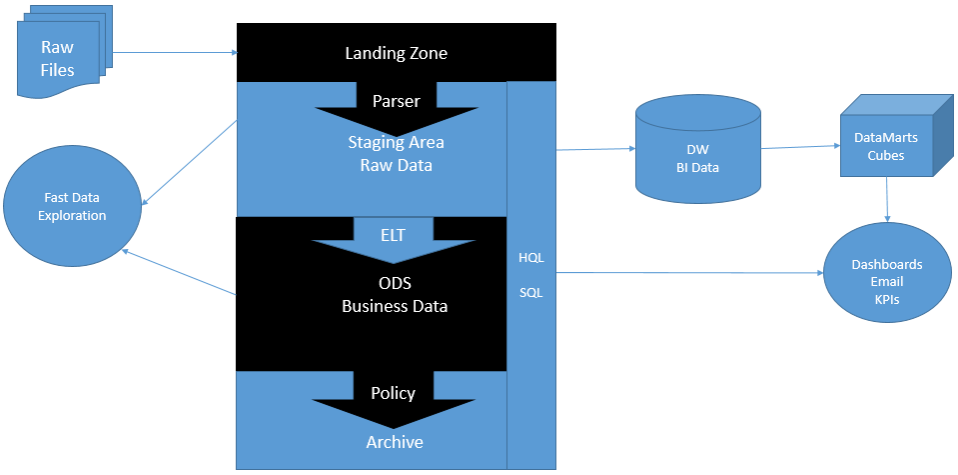
Most important Dimensions.



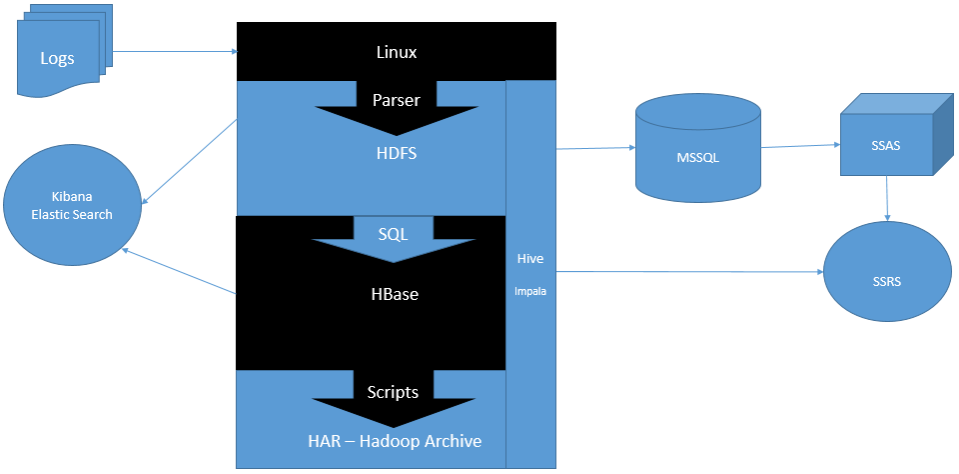

### Data Hub - Architecture View




**Data Hub - Functional View**




**Data Hub - Technical View**




## Data Layers

Having online access to data in its raw, source form — "full fidelity" data — means it will always be possible to perform new processing and analytics with the data as requirements change.

One size does not fit all: With this project, PSafe start to have a variety of different data storage technologies for different kinds of data. Our Implementation is compliant with Cloudera's recomendation.

In the RDBMS world, database features like scaling and replication are the hard parts left to the user. This worked fine in yesterday's enterprise when scale wasn't a big issue. Today it's quickly becoming the issue.

HBase was a natural choice, closely aligned with Hadoop and developed by the same community. HBase provides a record-based storage layer that enables fast, random reads and writes to data, complementing Hadoop by emphasizing high throughput at the expense of low-latency I/O.

**This is the biggest goal of the project, queryable warm and hot data. New data storage layers are:**

| Data Type | Volume 1 year (GB) | Usage | Data Layer | Storage | File Type | Archive | Retention | ETL - Input | Interface | Compression |
|---|---|---|---|---|---|---|---|---|---|---|
| Raw files | 38 TB | Cold Data | Landing Zone | HDFS | text | offline | 1 year | Java Parser | Map Reduce | gzip |
| Raw data | 19 TB | Warm Data | Staging Area | HDFS | Sequence FIle | HAR | 2 years | Java / HQL | Hive | Snappy |
| Business data - master data | 1 TB | Hot Data | ODS | Hbase | Hfile | Not required | Not required | SQL | Impala / Hive/ Apps | Snappy |
| BI Data | 0,2 TB | Hot Data | DW / DM | MSSQL / SSAS | mdf | Not required | Not required | Sqoop | T-SQL, MS BI tools | Not required |

OBS: Compressed files in Hive: SequenceFile + Snappy codec

## Data Patterns - ODS / HBAse - Requested Information

All actual reports must be supported. Please check ODS [http://wiki.psafe.com/doku.php?id=documentacao:bi:architecture#ods] section to see fact table layouts.

New BI Environment must provide these major business information:

#### Android User Profile - 360 Columns - Data Team Internal Demands

- Device Installation datetime
- Server Installation datetime
- ETL Insert datetime
- ETL Last Update datetime
- Android OS version
- Device model type
- Device manufacturer (brand)
- Device carrier
- Used Appbox in the first day installed? yes/no.
- Status (Active or Inactive). Also, Activation and Inactivation datetime
- Last Active Datetime - for Lifetime and retention
- 360 IP, Device IP, Session IP (for fraud [http://wiki.psafe.com/doku.php?id=documentacao:bi:architecture&#fraud_reports] detection)
- 360 Country, Device Country, Session Country (for fraud [http://wiki.psafe.com/doku.php?id=documentacao:bi:architecture&#fraud_reports] detection). Device country, set by user, also should be saved, but is not valid for fraud.

#### Android User Profile - MAT Columns - Data Team Internal Demands

All these cloumns are required from MAT for audit, ad-hoc and Retention actual Reports.

- created (installation date)

- IP
- Country
- Campaing
- Publisher
- Subpublisher (need many MAT columns to findout the subpublisher)
- Android OS version
- Device model type
- Device manufacturer (brand)
- Device carrier
- Payout
- google_id, device_id, os_id,
- ETL Insert datetime - MAT
- ETL Last Update datetime - MAT

#### Android User Profile - Features Usage Control - Product Team Demand

This demand only applies to major features. Listed columns below answers all questions.

- First datetime used
- Last datetime used
- Usage Count

Major features are:

| Feature ID | Function Name |
|---|---|
| 1000 | Opened app |
| 1001 | Verification |
| 1002 | Antivirus |
| 2001 | Limpeza |
| 1010 | Antitheft - entrou |
| 10000 | cofre - opened |
| 1300 | side bar (main) |
| 1003 | salvar dados |
| 1503 | monitorar dados |
| 1308 | bateria version 1.8 |
| 1310 | gerenciador de apps version 1.8 |
| 1006 | bloquear spam - entrou |
| 1203 | entrou Total Apps (APP TAB/MAIN Window) |
| 2002 | Bateria version 1.7 |
| 2004 | gerenciador de apps version 1.7 |
| 37000 | Verificação |
| 37001 | Cofre |
| 37002 | Limpeza |
| 37003 | Antivirus |
| 37004 | Antifurto |
| 17005 | Entered APPBOX from SMARTBOX |
| 7033 | Entered APPBOX from FLOATING WINDOW |
| 9998 | Entrar no appbox por um deeplink |
| 30002 | Clicar no smartbox tipo 1 |
| 30004 | Clicar no smartbox tipo 2 |
| 33002 | Clicar no último item dos apps sugeridos na float window |
| 33003 | Clicar no botão da home screen |
| 33005 | Clicar em push notification com deeplink do kahuna |
| 17001 | Entered TOP CHART tab |
| 17002 | Entered FEATURED tab |
| 17003 | Entered APPS tab |
| 17004 | Entered GAMES tab |
| 1203 | Entered APPBOX from APP TAB |

### Data Patterns - ODS / Hive - Requested Information

All actual reports must be supported. Please check ODS [http://wiki.psafe.com/doku.php?id=documentacao:bi:architecture#ods] section to see fact table layouts.

New BI Environment must provide these major business information:

#### AppBox User Profile - Data Team Demands

The key table in Hive will be:

- IMEI
- APK
- Time/Date/MONTH
- Campaign
- Event

#### AppBox User Profile - Product Team Demands

We will have, from mobile_kill_Appinfo and Appbox Filtering, all installed Apps in the device. So we can report about:

- Already installed Apps
- Already ignored Apps (many impressions, none installation). Maybe just sum all impressions per APK per IMEI.
- Already Deleted Apps

**Objective: REAL TIME RECOMENDATIONS PER USER, REAL TIME CAMPAIGN OPTIMIZATION.**

#### Reports - New Demands

## Reports - New Demands

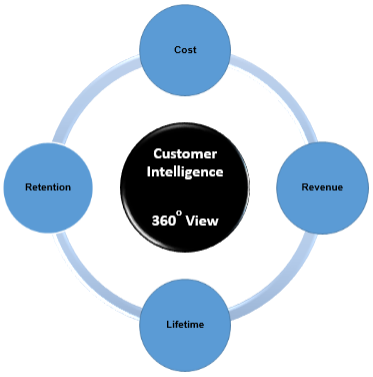
This are the major new demands.

### Customer Intelligence - 360 degrees Report

This report cares about Cost, Retention, lifetime and revenue.

Required information per IMEI:

- Cost - MAT payout
- Retention - Days, publisher
- Lifetime - Days, Campaign
- Revenue - sum of: a) installed apps payd by installation. b) Open apps payd by AppOpen)




### Android Reports - New Users Demands

- Overlap: Date, product, qtd
- Feature use of our Monthly Active Users - new fact table.
- How many unique users have actively engaged with the app, either opened the app or did something with the notification bar or the float window - new fact table.
- TOP 5000 Apps: date, pkg name, total. This data will be pushed in MSSQL, so we also need insert_Datetime, for tracking.
- Profile Analysis. Breakdown of our users (all users, MAUs) - language, location (country is fine for now), version of app, version of Android, device, screen resolution - Data Export.
- Add Total QTD to FactAnFeatures table at DW. Today this table QTD column is unique value.
- FUTURE, NO ETA…2015: Features entrance and exit datetimes. User flow trought the APP.

### Appbox Reports

Ad Hoc reports, to be automatized:

- Unique Installs per APK - (client_date, server_date,user_id (imei),type, language, country, version, machine, publisher)
- Avg downloads and installations per unique user. Per day.
- Number of impressions to generate install. By App.
- The number of AppBox DAUs that came from new installations by day
- The number of AppBox DAUs that never used Appbox before.
- AppBox WAUs – Last 4 weeks by week
- AppBox MAU
- Frequency distribution of total apps installs in the last 30 days by device for users who have visited the app box – how many devices have installed 0 app, 1 app, 2 apps etc

### Facebook Extraction

A file with android_id (only) must be sent to Ram's team in 4 diffenrent files:

- Brazil active users
- Brazil inactive users
- NonBrazil active users
- NonBrazil inactive users

OBS: Facebook support these 3 ids: Android's advertising ID OR Apple's Advertising Identifier(IDFA) OR Facebook App User IDs.

### Fraud Reports

MAT installations outside campaign country. All of these are Ram's requests….

- Frequency distribution by the sub-net: Publisher, subpublisher, subnet (3/4 of IP), total. Criteria is: 360 country is NOT BRAZIL, MAT country is BRAZIL and IP COUNTRY is BRAZIL (Bad data)
- Fraud Auditing: If there is a difference in our internal tracked country and what MAT has
- Revenue Auditing: Looking a segments to assist in solving revenue yield problems with affiliates
- Targeting: Yield for Apps in our app box WILL vary tremendously by this data set, which can be used to drive targeting / optimization

### Total Windows Reports

- Bitdefender: Total Installations (active subscriptions), Uninstallations ⇐ 14 days (cancelled trial offer), Uninstallations > 14 days (cancelled subscription), Active PCs (running bitdeffender).

## Data Sources

Layouts of the most important data sources (1st wave)

**APPBOX**

| Param | Description |
|---|---|
| k1=imei | MD5 of IMEI |
| K2=os | Always be "Android" |
| K5=version | Version of psafe,for example,1.5.0.1014 |
| K6=machine | Including model,manufacture,system version, code of sim card provider and MCC,separated by "%" |
| K9=data | AppBox statistic log data(json array):<br>mType: log type, 1means download(click), 2 means install, 3 means show.<br>mTimeStamp: log timestamp<br>mLine: line in the list<br>mPackName: package name |
| K10=uiversion | Always be 20003 |
| K11=Channel | Publisher id |
| K12=Language | Two letter code of language |
| K13=Country | Two letter code of country |
| K19=android id | Android ID |

**ANDROID OPERATIONAL**

| Parameter | Description |
|---|---|
| imei | MD5 of IMEI |

| | |
|---|---|
| os | Always be "Android" |
| version | Version of PSafe Total, for example,1.2.0.1038 |
| machine | Including model, manufacture, system version, code of sim card provider and MCC, separated by "," |
| data | Feature usage. "%1000,3" means feature id:1000 were used 3 times |
| uiversion | Always be 20003 |
| Channel | Channel id |
| Language | Language abbreviation |
| Country | Country abbreviation |
| android id | Android ID |
| gp_referrer | Dynamic parameter of Google Play Market url, reserved. |

Last Layouts version - complete list: psafe.logs.20141008.docx

### Data Flow

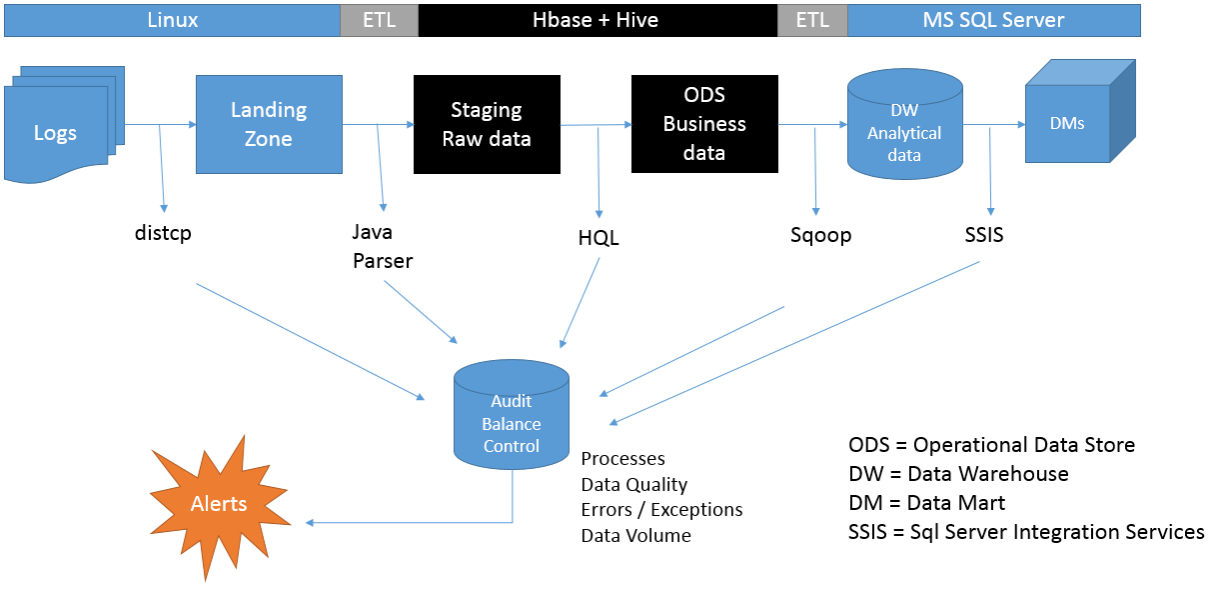
Data only moves in one direction. So, ODS does not receive data from DW, as is in currect state.

Files bigger than 100 MB and compressed with gzip (does not split) are uncompressed and processed with map reduce jobs. Cloud Scan for example. Files smaller than 100 MB does not parallel processing, so they are processed as is.

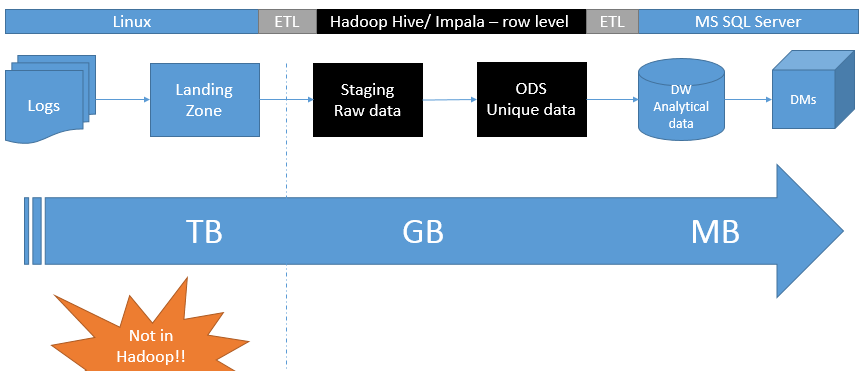
Pure Java will be replaced by PIG_JAva plugins in a 3rd wave.

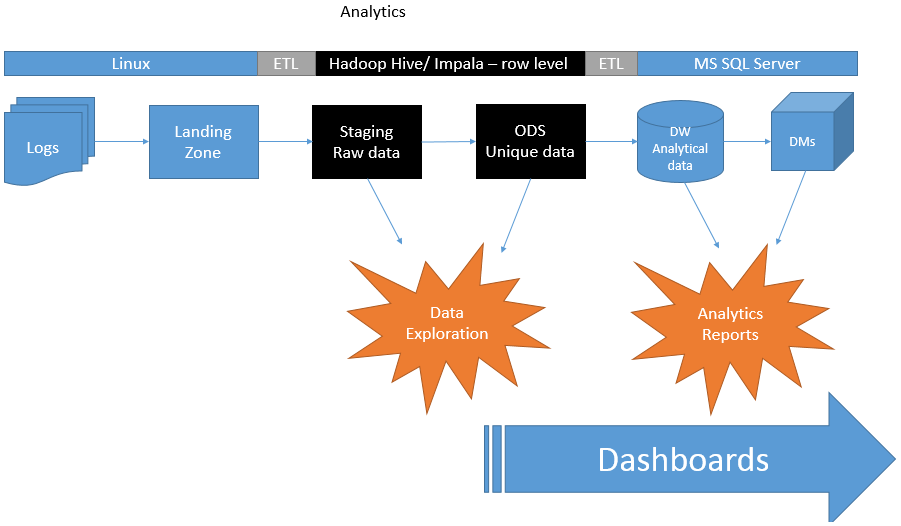
Data will be exported from ODS to DW using Sqoop. All nodes need access to MSSQL.

Basic Data Flow




Data Volume




**Sqoop**

Sqool exports are performed by multiple writers in parallel. Each writer uses a separate connection to the database; these have separate transactions from one another. Sqoop uses the multi-row INSERT syntax to insert up to 100 records per statement. Every 100 statements, the current transaction within a writer task is committed, causing a commit every 10,000 rows. This ensures that transaction buffers do not grow without bound, and cause out-of-memory conditions. Therefore, an export is not an atomic process. Partial results from the export will become visible before the export is complete.

For this Sqoop process, we will create staging tables within ODS and DW layers:

- Same strucuture as Fact tables.
- Truncated every day after data movement, in both ODS and DW layers.
- business key instead of dimensions keys.
- From staging tables in DW to Fact tables, business keys will be replaced with Dimensions keys. This will be done with SSIS and T-SQL.

## Data Quality

Our process to Data Quality are:

**ABC - Audit, Balance and Control**

Data auditing is the process of doing a profile check and assessing the quality of data, to find how accurate it is. This can be achieved by keeping track of all the data changes. We need a queryable table at Hive with:

- ETL name
- start time
- end time
- status
- Error message
- rows added, updated and rejected
- Processed file name, size and location.

**Alerts**

- Failed Jobs (list)
- Rejected rows (total per data source)
- BPM: Warning for KPIs variation above 15% (suggested value, to be certified).
- Missing files: control expected number e total size of files per data source.

**Corrupted files**

- Must be moved to corrupted files folder.
- ETL process can't stop.
- Problem must be registered at ABC looging base: file name, size, date time, error message.

**Differences**

Differences between 2 clusters are expected:

- We have different validations (new converter job which is different from the old one).
- As a result of different validations we have more input for mobile_operational_data (e.g. for one of the days we had 18.1MM of input lines after conversion on old cluster and 18.4MM of input lines in a new one).
- Processing is different (before we had pure map reduce and now we use Hive).
- We have new dimensions. Our structure has changed and we have new dimensions, just like in staging_db.Exp_FactAnActivePCs.

**Rejected Records (queryable table at Hive)**

- Rejected data is kept on Staging Area, for debuging, market as rejected.
- Simple rule to reject data is: Reject any incomplete log record. All layout columns are mandatory. Exception is Appbox, list and campaign may not come.
- Problem must be registered at ABC looging base: file name, size, date time, error message.

## Dimensions

| Data Sources | | Dimensions | | | | | | | | | Reference Data | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Product | Folder | Id | Channel | Language | version | OS | Manufacturer, Model | Country / Locatin | APPs | Campaign | Features / Events | Overlap | Installation | Uninstall |
| Psafe Total Windows | pc_operational_data/s.psafe.com | mid | x | x | x | x | | | | | x | no | x | reason |
| Psafe Total Windows | pc_total_startupitems | mid | | x | | | | | | | | yes | | |
| Psafe Total Windows | crash_pctotal | mid | | | x | x | ie | | | | | no | | |
| Psafe Total Windows | security_service_data | mid | | x | x | x | ie | | | | | yes | | |
| Psafe Internet | pc_operational_data/s.psafe.com | mid | x | x | x | x | | x | | | | no | x | |
| Psafe Total Mac | pc_operational_data/s.psafe.com | mid | x | x | x | | | | | | x | no | x | status |
| Psafe Total Android | mobile_operational_data | imei, os_id | x | x | x | x | x | x | | | x | no | x | |
| Psafe Total Android | mobile_Appbox | imei, os_id | publisher | x | x | x | x | x | x | x | x | yes | x | |
| Psafe Total Android | crash_mobiletotal | imei | | | x | x | | | | | x | no | | |
| Psafe Total Android | security_service_data/*Mobilekill* | imei | x | x | x | x | x | ip | x | | x | yes | | |
| Psafe Total Android | security_service_data /*Mobilekill_register* | imei | x | x | x | x | x | ip | x | | x | yes | | |
| Psafe Total Android | security_service_data /*Mobilekill_reqstat* | imei | x | x | x | x | x | ip | x | | x | yes | | |
| MAT | MAT | google id | publisher | x | x | x | | x | | x | | no | x | |
| Psafe Total Android | security_service_data_mobile | | | | | | | | | | | | | |

Other dimensions: Date, Month, Hour.

## KPIs

| | Product | | | | |
|---|---|---|---|---|---|
| KPI | Appbox | Psafe Internet | Total Android | Total Mac | Total Windows |
| AC-U | x | x | x | x | x |
| AC_DAU | x | x | x | x | x |
| DNI | x | x | x | x | x |
| DAU (Total, Country, Language, Version) | x | x | x | x | x |
| MAU (Total, Country, Language, Version) | x | x | x | x | x |
| WAU (Total, Country, Language, Version) | x | x | x | x | x |
| WAU and MAU by feature | | | x | | |
| Revenue per Country, Version, Campaign (Install and AppOpen) | x | | | | |
| Bitdefender: Uninstall <- 14 days, uninstall > 14 days, active Bitdefender users, Total Uninstallations | | | | | x |
| eCPC, eCPM, Install Rate, CTR, Unique Impressions, Unique Downloads, Daily New App Opens, App Open Rate | x | | | | |
| Retention per Publisher | | | x | | |
| Retention per SubPublisher | | | x | | |
| Retention per Country | | | x | | |
| AVG downloaded APKs | x | | | | |
| Lifetime per Publisher | | | x | | |
| Lifetime per Subpublisher | | | x | | |
| Lifetime per Country | | | x | | |
| Gross M. A - LTV Analysis by Network | | | x | | |
| Gross M. A - Publisher, Media Margin by Network/Publisher | | | x | | |
| DEU, WEU, MEU | | | x | | |

| |
|---|
| 1st Wave |
| 2nd Wave |
| 3rd Wave |

Please check the Data Dictionary for definitions and business rules.

## KPIs Delivery Calendar

| | Month | | |
|---|---|---|---|
| Project Delivery | nov/14 | dez/14 | jan/15 |
| 1st Wave | x | x | |
| 2nd Wave | | x | x |
| 3rd Wave | | x | x |

## ODS

Operational Data Storage contains Business data, what is raw data after business rules:

- HBase - Master data, user profile tables at HBase.
- HDFS/Hive - Operational data, Fact tables ready to feed temp tables in DW with Sqoop. Please check attached script of destination tables. We can't have all data into Hbase because on known performance issues.

dw_fact_tables.docx

## ODS - User Profile at HBase

### HBase

**PSafe Total Android user profile schema**

---

**Data Access Patterns**

Write:
1. Create profile.
2. Update profile.

Read:
1. Select profiles by time range (DNI, DAU, WAU, MAU, retention, etc).
2. Aggregate profiles by dimensions (channel, version, country, etc).
3. Select user packages installed.

Android:

| TABLE: User Profile | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Column Family: profile | | | | | | | | | | | | | | | Column Family: activity | | | |
| | version | channel | first_install | android_id | machine | ip address | ui version | language | country | gp referrer | mat publisher | mat sub publisher | last active | last upgrade | | timestamp 1 | timestamp ... | timestamp N | |

| Row Key: user_id (IMEI md5 / Google AID) | string | string | long | string | string | string | string | string | string | string | int | int | long | long | int[] | int[] | int[] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | v14 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | v14 | v0 | v0 | v0 |
| | v... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | v... | | | |
| | v0 | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | v0 | | | |

## HBase

### PSafe AppBox user profile schema

---

**Data Access Patterns**

Write:
1. Create profile.
2. Update profile.

Read:
1. Select events (install, show, click) by time range.
2. Select user packages installed by time range.

AppBox

| TABLE: User AppBox | | | | | |
|---|---|---|---|---|---|
| Row Key: user_id (IMEI md5 / Google AID) + package | Column Family: appbox events | | | Column Family: overlap | |
| | event + timestamp | event + timestamp | event + timestamp | install | uninstall |
| | byte[] - log data in protobuf (ch, os, lang ...) | ... | byte[] - log data in protobuf (ch, os, lang ...) | long | long |
| | v0 | ... | v0 | v3 | v3 |
| | | | | ... | ... |
| | | | | v0 | v0 |

ODS Scripts - Tables:

script_hive_ods.docx

## Risks

These are the biggest risks of the project:

**Disconnected Approach**

Not integrated data creates:

- Duplicative data across the enterprise
- An inability to deploy truly effective analytics
- Significantly higher costs for IT

**How to Mitigate**

The easiest way to tear down these walls and connect the data islands is to make the commitment up front that the Data Architecture will include all data meaningful to the organization so that design and management can be aligned with enterprise goals and expectations. Please refer to the Data Sources [http://wiki.psafe.com/doku.php?id=documentacao:bi:architecture&#data_sources] section.

**Lack of Actionable Governance**

Governance tends to be both a workflow process as well as an associated toolset. If either the tool or the workflow process has limitations or is not present, the Governance effort generally fails.

**How to Mitigate**

List, analyze and develop:

- Integration of Data Governance with all pertinent system management processes.
- Integration of Data Governance with security standards and processes.
- Integration of Data Governance and Portfolio Management processes.
- Integration of Data Governance and Data Architecture.
- Adherence to clearly defined metrics, tracked on a continual basis. (this can include ALM metrics, integrity metrics as well as performance metrics).
- The Governance process has clearly defined roles with authority and decision gates built in.
- The Governance is closely aligned to both tactical and strategic organizational goals.

**Limited View of Future Demands**

If your project is narrowly focused on deployment of one particular product or technology, then it is obvious that the product architecture will be the primary focus of the effort.

**How to Mitigate**

Having online access to data in its raw, source form — "full fidelity" data — means it will always be possible to perform new processing and analytics with the data as requirements change. Also, open architecture and schema on read solutions speed up any new requirement. Please refer to Architecture Drivers [http://wiki.psafe.com/doku.php?id=documentacao:bi:architecture#architecture_drivers] section.

**Lack of Data Relationships and business Semantic**

Big data technology, often as not, obscures the fundamental semantics of business relationships. Which is kind of ironic, because big data is supposed to "reveal", not "hide". The base of this issue is the lack of schema.

**How to Mitigate**

Data documentation. Please refer to raw data [http://wiki.psafe.com/doku.php?id=documentacao:bi:360_raw_log_files] section.

# Reference Architectures

## Cloudera


## Sensitive Information

Customer data protection is a corporate social responsibility. Businesses must start protecting customer privacy and data as conscientiously as they protect their own. SI is a sub-set of personal information and demands higher level of protection. Australian Governament leads the world about this matter and lists SI examples:

- racial or ethnic origin;
- political opinions;
- membership of a political association;
- religious beliefs or affiliations;
- philosophical beliefs;
- membership of a professional or trade association;
- membership of a trade union;
- sexual preferences or practices;
- criminal record;

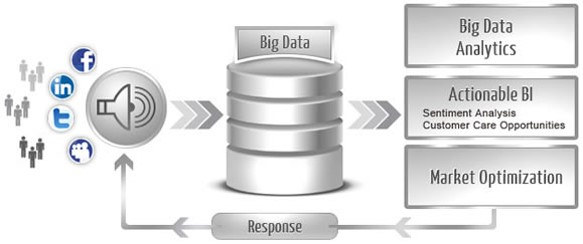
**PSafe has no SI from our users or clients.**

Ref:http://www.alrc.gov.au/publications/6.%20The%20Privacy%20Act%3A%20Some%20Important%20Definitions/sensitive-information [http://www.alrc.gov.au/publications/6.%20The%20Privacy%20Act%3A%20Some%20Important%20Definitions/sensitive-information]

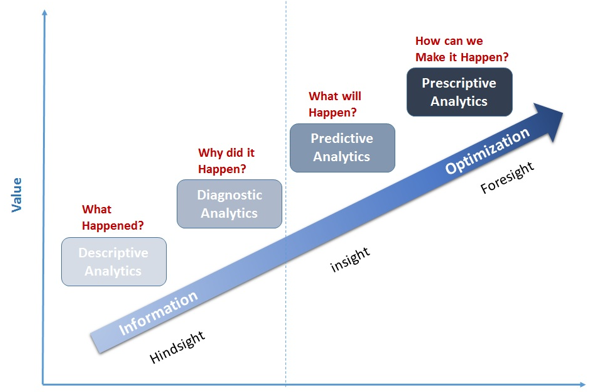
## TO BE IMPLEMENTED

Second generation architecture…

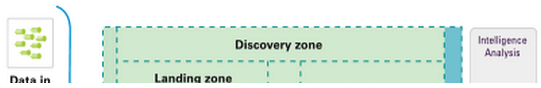
- PIG combined with Java (plugins).
- Automated data profiling and relationship discovery
- Use Impala do query HBase historical data.
- Lostash in the Statistics Servers so we can replace java parser and use Elastic Search for monitoring.
- In Memory Data Grid.
- Zabbix for monitoring.
- Automatic Business Response.
- Android User Profile updated in real time at Hbase.
- Full reaload of UsersAndroid table at MSSQL.
- BPM: KPIs expected values, business performance x estimations, alerting dashboards.
- Fast Data Exploration.
- Real Time user profiles for Appbox listed apps.
- Conceptual Understanding of the data: Google IMO and uninstall reason.
- Social Intelligence.

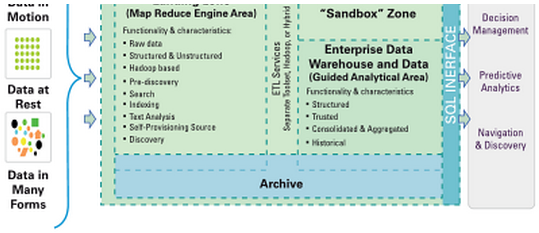



- Predictive Analytics / Data Mining / Advanced Analytics / Data Insights.




- Discovery Zone: A virtual semantic layer to real data discovery.






Data discovery is becoming an increasingly important activity for organizations that rely on their data to be a differentiator. Today, that describes most businesses, as the ability to see trends and extract meaning from available data sets applies to almost any industry.

What this requires is two critical components: analysts with the creativity to think of novel ways of analyzing data sets to ask new questions (often these kinds of analysts are called data scientists); and to provide these analysts with access to as much data as possible.

---

/var/www/html/dokuwiki/data/pages/documentacao/bi/architecture.txt · Última modificação: 15/05/2015 18:26 por Sergio Bohrer