Trabajando con

APACHE ATLAS

BIG DATA – BUSINESS INTELLIGENCE – MACHINE LEARNING

CONTENIDO

1. INTRODUCCIÓN

Apache Atlas es una herramienta *open-source*, con licencia Apache 2.0, para la gobernanza del dato la cual permite la integración con todo el ecosistema de datos de las empresas.

Atlas permite crear un almacén de metadatos centralizado para saber dónde encontrar un conjunto de datos dentro de la empresa (**Data Discovery**), permite saber qué cambios ha sufrido y que transformaciones se le han realizado a los datos originales a lo largo del tiempo (**Data Lineage**) y centraliza la seguridad para saber quién puede acceder a esos datos y quién puede modificarlos (**Data Security**) apoyándose en Apache Ranger.

Pricing

Actualmente está integrada dentro de la plataforma de Cloudera, la cual ofrece además una versión Entreprise que añade soporte por parte de la empresa. Su pricing es el siguiente:

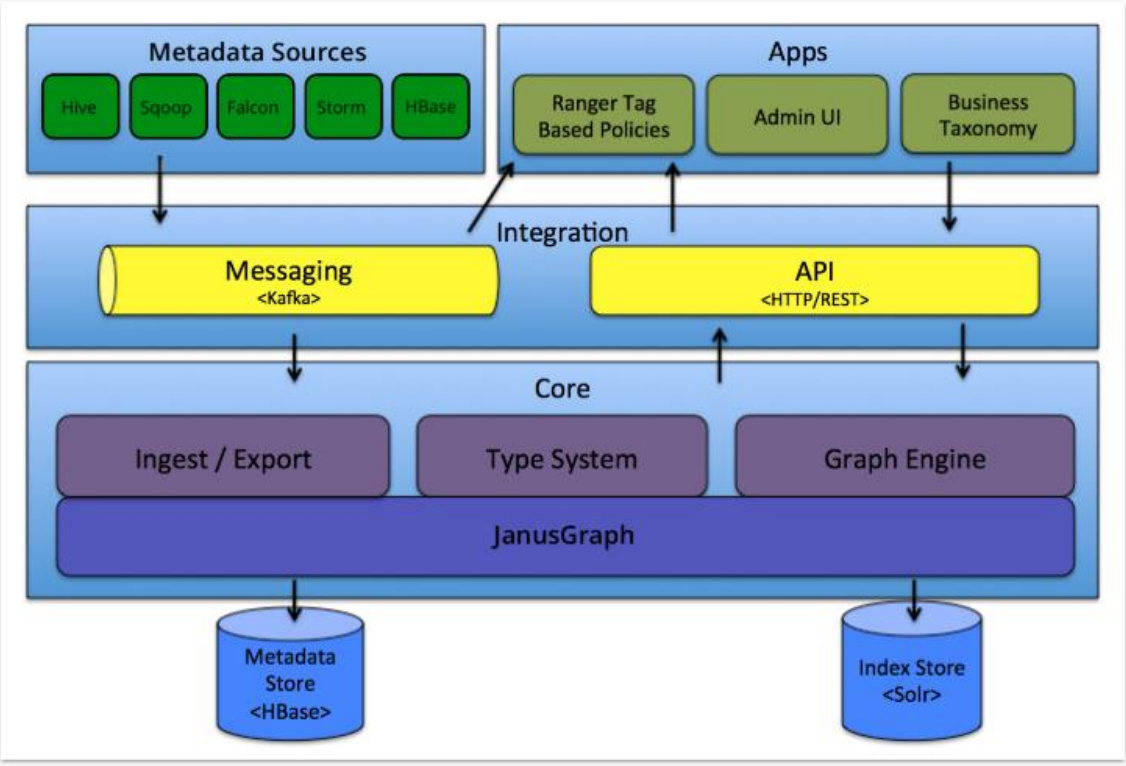
Platform pricing

Cloudera Platforms Annual subscription pricing	CDP Data Center	Enterprise Data Hub	HDP Enterprise Plus
Base price Per node	$10,000 [1]	$10,000	$10,000
Variable price - compute [2] Per CCU over 16 core, 128 GB node cap	$75 [3]	$75 [4]	$75 [4]
Variable price - storage [2] Per TB over 48 TB node cap	$25 [3]	$25 [4]	$25 [4]
Conversion entitlement [5] For CDP Data Center	-	Yes	Yes

Se puede obtener una comparativa entre las distintas versiones a través de este enlace.

Arquitectura

La arquitectura de Apache Atlas y las herramientas en las que se apoya son las siguientes:



Como fuente de metadatos puede utilizar: HBase, Hive, Sqoop, Storm, Kafka y Falcon (en desuso).

Instalación y despliegue

Apache Atlas se puede instalar a través de la plataforma de **Apache Ambari** que se puede instalar en los siguientes sistemas operativos: RHEL/CentOS/Oracle Linux 7, SLES 12, Ubuntu 16, Ubuntu 18 y Debian 9.

Los **prerrequisitos** para su instalación son: HBase, Kafka y SolrCloud/Elasticsearch

Atlas se ha diseñado para desplegarse como **servidor**, pudiendo conectarse a ella desde un navegador.

2. DATA DISCOVERY: ALMACÉN DE METADATOS

Atlas permite a los usuarios definir un modelo para los objetos de metadatos que desean administrar. Todos los objetos de metadatos administrados por Atlas que están listos para ser utilizados se modelan usando "Tipos" y se representan como "Entidades". Antes de empezar a utilizar Atlas es importante entender estos conceptos.

Para almacenar y acceder a un tipo particular de objetos de metadatos, Atlas, permite definir "**Tipos**". Similar a las "clases" de Java o al "esquema de una tabla" en base de datos relacionales.

An example of a type that comes natively defined with Atlas is a Hive table. A Hive table is defined with these attributes:

```
Name:         hive_table
TypeCategory: Entity
SuperTypes:   DataSet
Attributes:
    name:             string
    db:               hive_db
    owner:            string
    createTime:       date
    lastAccessTime:   date
    comment:          string
    retention:        int
    sd:               hive_storagedesc
    partitionKeys:    array<hive_column>
    aliases:          array<string>
    columns:          array<hive_column>
    parameters:       map<string,string>
    viewOriginalText: string
    viewExpandedText: string
    tableType:        string
    temporary:        boolean
```

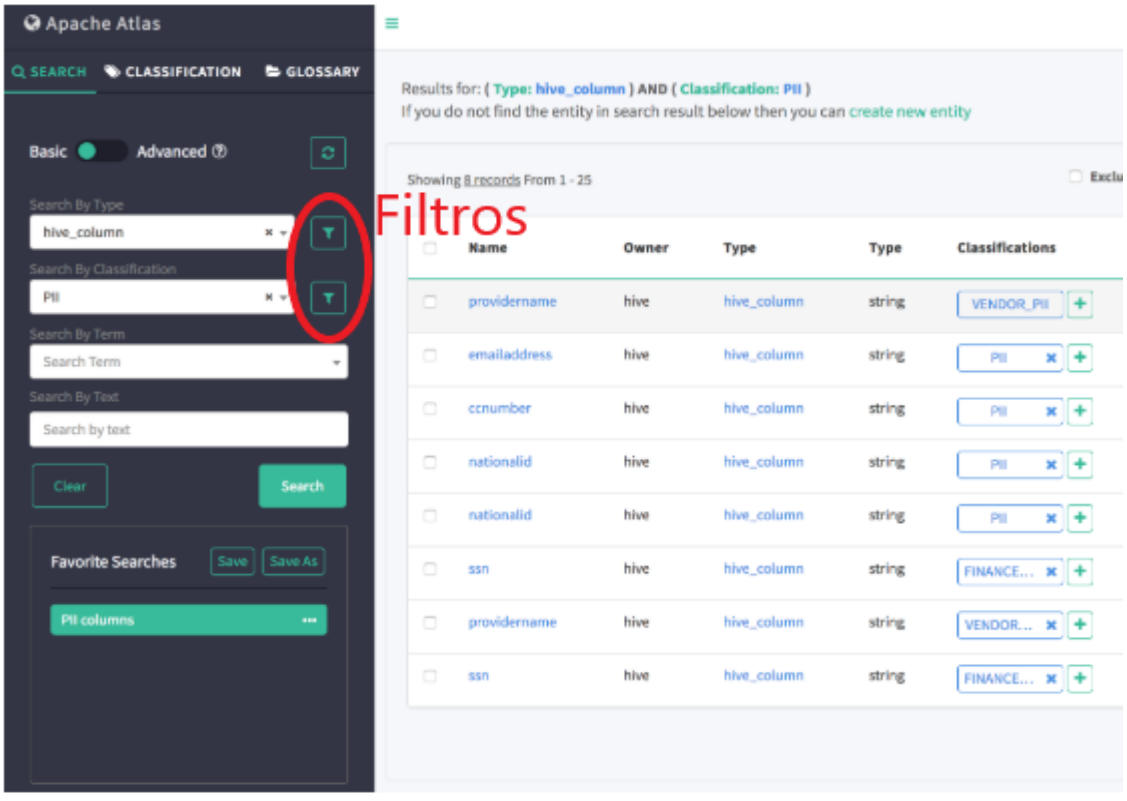
Por cada "Tipo" creado en Atlas se pueden crear diferentes instancias o "**Entidades**". Similar a la "instanciación de objetos de una clase" en Java.

An example of an entity will be a specific Hive Table. Say Hive has a table called 'customers' in the 'default' database. This table will be an 'entity' in Atlas of type hive_table. By virtue of being an instance of an entity type, it will have values for every attribute that are a part of the Hive table 'type', such as:

```
guid:     "9ba387dd-fa76-429c-b791-ffc338d3c91f"
typeName: "hive_table"
status:   "ACTIVE"
values:
    name:            "customers"
    db:              { "guid": "b42c6cfc-c1e7-42fd-a9e6-890e0adf33bc", "typeName": "hive_db" }
    owner:           "admin"
    createTime:      1490761686029
    updateTime:      1516298102877
    comment:         null
    retention:       0
    sd:              { "guid": "ff58025f-6854-4195-9f75-3a3058dd8dcf", "typeName": "hive_storagedesc" }
    partitionKeys:   null
    aliases:         null
    columns:         [ { "guid": ""65e2204f-6a23-4130-934a-9679af6a211f", "typeName": "hive_column" }, { "guid": ""d726de70-faca-
46fb-9c99-cf04f6b579a6", "typeName": "hive_column" }, ...]
    parameters:      { "transient_lastDdlTime": "1466403208"}
    viewOriginalText: null
    viewExpandedText: null
    tableType:       "MANAGED_TABLE"
    temporary:       false
```

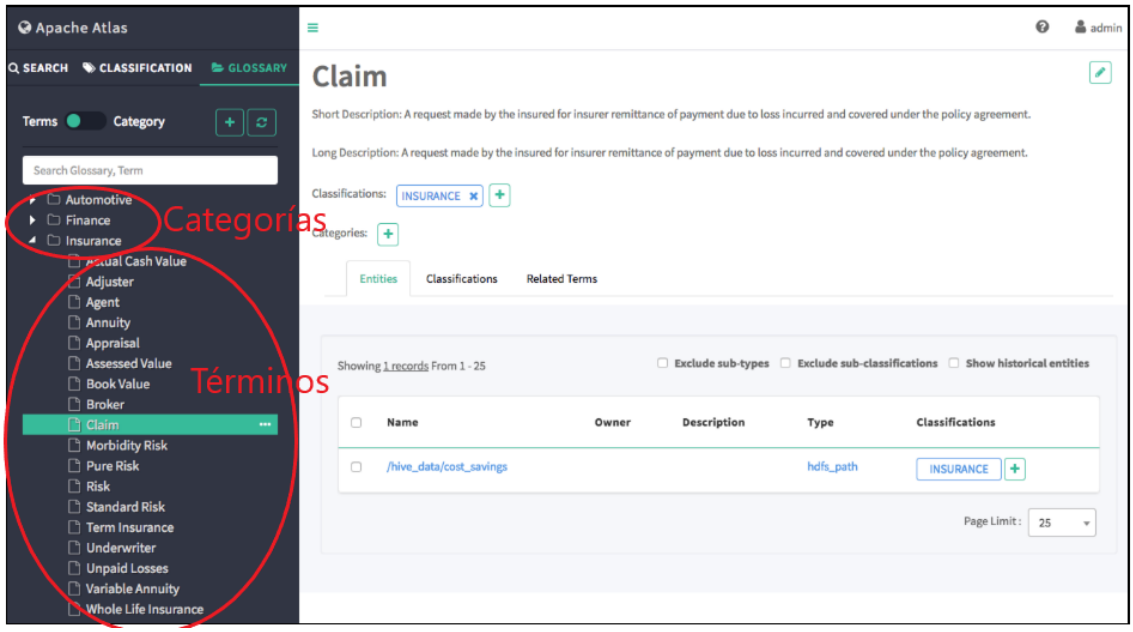
Interfaz de usuario

Para empezar a utilizar datos se pueden buscar y encontrar fácilmente con su buscador ya sea de manera básica utilizando texto y apoyándose en filtros, o de forma más avanzada mediante DSL. En la siguiente captura se puede observar como sería el buscador básico:



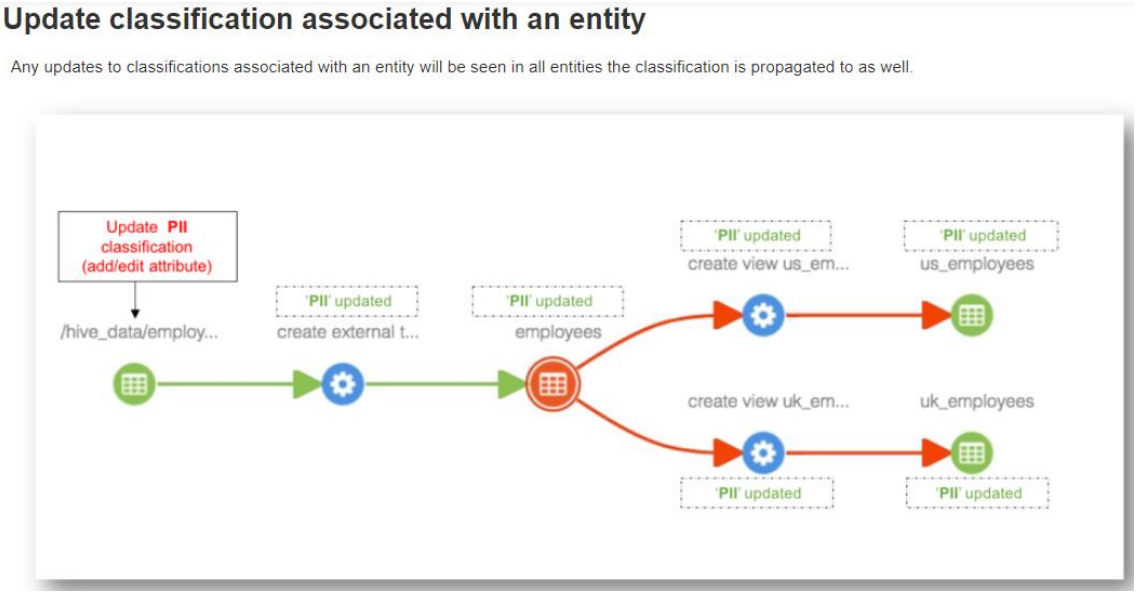
Glosarios

Para facilitar la utilización de la herramienta a los usuarios de negocios Atlas permite crear Glosarios donde se puede relacionar términos y categorías con los datos para que sean más fácil de encontrar. En la siguiente captura se puede ver una clasificación de distintos datos mediante categorías y términos:



3. DATA LINEAGE: SEGUIMIENTO DE TRANSFORMACIONES

Apache Atlas permite hacer un **seguimiento de las transformaciones** que han sufrido los datos originales y en que datos derivados han terminado para poder tener un control total sobre ellos. En esta captura se puede observar un caso de uso:

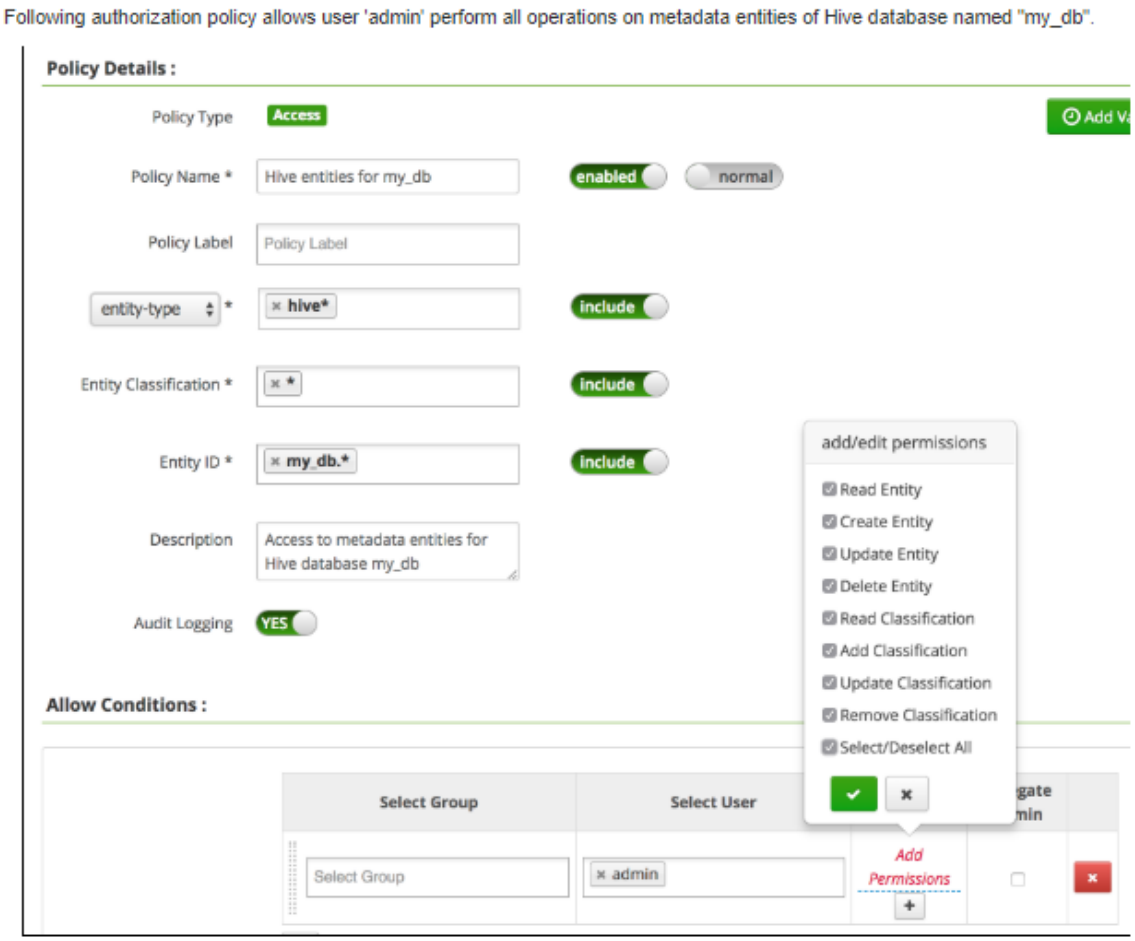


Para ver más ejemplos se puede acceder a este enlace.

4. DATA SECURITY: AUTENTIFICACIÓN Y PERMISOS

Atlas permite la autentificación de personal con los siguientes métodos: **LDAP**, **File** y **Kerberos**. Se pueden utilizar varios a la vez y en caso de que falle el que se ha establecido como método principal pasará a utilizar otro como escenario alternativo.

Además, permite la creación de usuarios con distintos **permisos** de forma básica mediante Java reg-ex y a estos usuarios se les puede modificar la política de autorizaciones de forma avanzada apoyándose en Apache Ranger. En la siguiente captura se puede ver un ejemplo del editor avanzado:



5. CONCLUSIONES

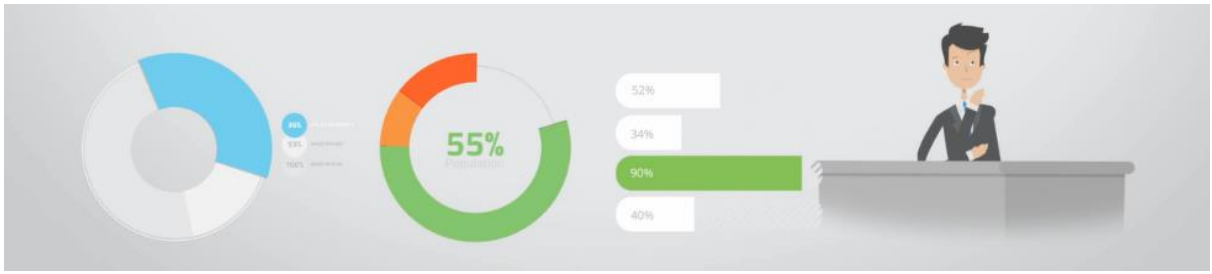
Apache Atlas es una herramienta eficaz para implementar el Gobierno del Dato ya que abarca los múltiples campos que se requieren para una buena gobernanza de dato, la herramienta es *open-source* y está diseñada para ser extensible y añadir funcionalidades a través de una API REST para así poder ofrecer una solución global. La empresa Comcast, por ejemplo, ha añadido a Atlas la posibilidad de añadir esquemas de Apache Avro, lo que le permite manejar de manera más eficiente la calidad del dato.

Aunque eficaz, desplegarla y utilizarla puede ser una tarea compleja ya que requiere de gran número de herramientas para que funcione correctamente (Apache Ambari, Apache Ranger, Kafka, HBase…), por lo que puede terminar en no ser una solución muy eficiente.

6. SOBRE STRATEBI

En Stratebi ofrecemos **gran cantidad de soluciones analíticas** por una compañía de **rápido crecimiento,** innovando en las áreas tecnológicas de mayor desarrollo en la actualidad: **Business Intelligence, Big Data y Social Intelligence**, muchas de ellas, basadas en soluciones **Open Source**.

Además, somos **Partners Certificados en Microsoft PowerBI, Talend, Kylin, Snowflake, Pentaho y Vertica**, con gran número de proyectos con ámbas tecnologías



Stratebi es una empresa española, con sede en Madrid y oficinas en Barcelona, Alicante y Sevilla, creada por un grupo de profesionales con amplia experiencia en sistemas de información, soluciones tecnológicas y procesos relacionados con soluciones de Open Source y de inteligencia de Negocio.

Esta experiencia, adquirida durante la participación en proyectos estratégicos en compañías de reconocido prestigio a nivel internacional, se ha puesto a disposición de nuestros clientes a través de Stratebi.

Stratebi es la única empresa española que ha estado presente todos los Pentaho Developers celebrados en Europa (Mainz-Alemania, Barcelona, Lisboa, Roma, Amsterdam, Sintra, Amberes (2) Londres…), habiendo organizado el de Barcelona.

En Stratebi nos planteamos como **objetivo** dotar a las compañías e instituciones, de herramientas escalables y adaptadas a sus necesidades, que conformen una estrategia Business Intelligence capaz de rentabilizar la información disponible. Para ello, nos basamos en el desarrollo de soluciones de Inteligencia de Negocio, mediante tecnología Open Source.

Stratebi **son** profesores y responsables de proyectos **del Master en Business Intelligence de la Universidad UOC, UCAM, EOI…**

Los profesionales de Stratebi son los creadores y autores del primer weblog en español sobre el mundo del Business Intelligence, Data Warehouse, CRM, Dashboards, Scorecard y Open Source.

Stratebi es partner de las principales soluciones Analytics: Microsoft PowerBI, Talend, Pentaho, Vertica, Snowflake, Kyligence, Cloudera…

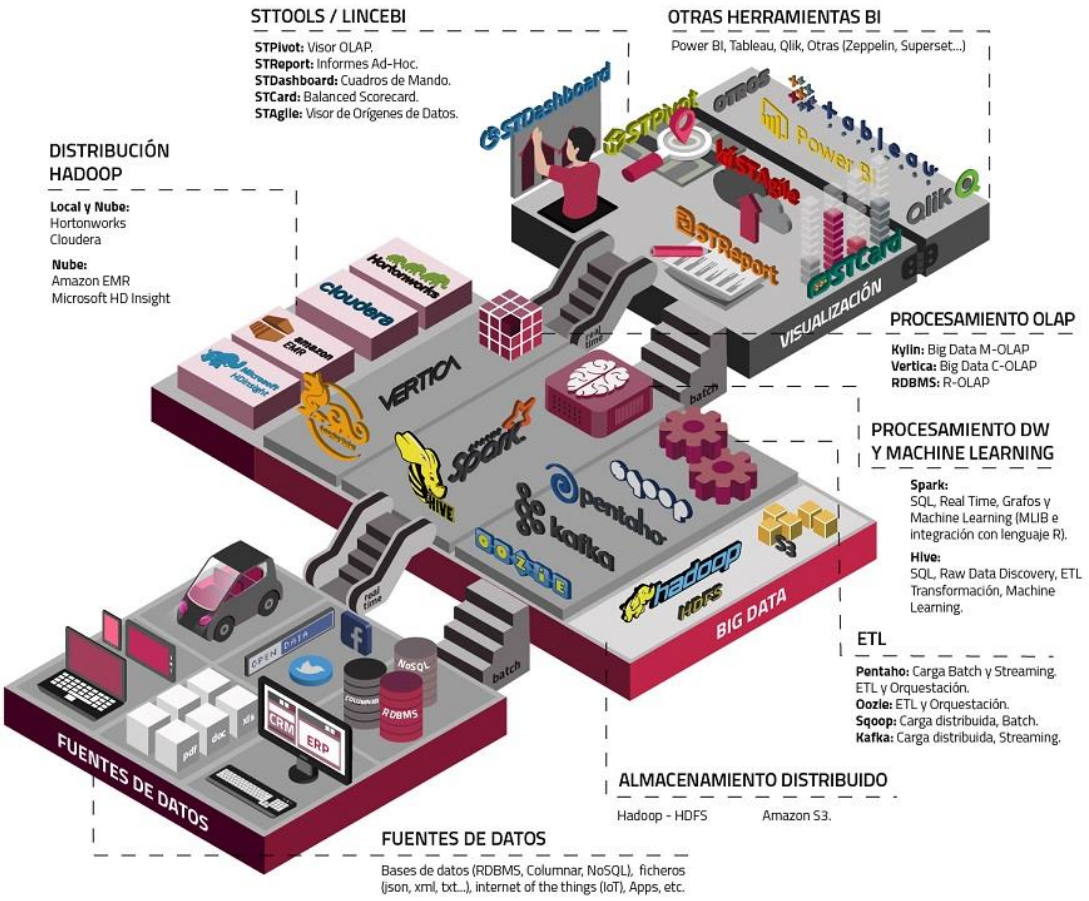
Todo Bi, se ha convertido en una referencia para el conocimiento y divulgación del Business Intelligence en español.

Desarrollamos nuevas soluciones analíticas basadas en Open Source, para la generación de Cuadros de Mando en tiempo real, con tecnologías IoT para SmartCities, machine learning, etc...

LinceBI
The best open source BI based solution

7. TECNOLOGÍAS

Recientemente, hemos sido nombrados Partners Certificados de Vertica, Talend, Microsoft, Snowflake, Kylligence, Pentaho, etc…


STTOOLS / LINCEBI
STPivot: Visor OLAP.
STReport: Informes Ad-Hoc.
STDashboard: Cuadros de Mando.
STCard: Balanced Scorecard.
STAgile: Visor de Orígenes de Datos.
OTRAS HERRAMIENTAS BI
Power BI, Tableau, Qlik, Otras (Zeppelin, Superset...)
DISTRIBUCIÓN HADOOP
Local y Nube:
Hortonworks
Cloudera
Nube:
Amazon EMR
Microsoft HD Insight
VISUALIZACIÓN
PROCESAMIENTO OLAP
Kylin: Big Data M-OLAP
Vertica: Big Data C-OLAP
RDBMS: R-OLAP
PROCESAMIENTO DW Y MACHINE LEARNING
Spark:
SQL, Real Time, Grafos y Machine Learning (MLIB e integración con lenguaje R).
Hive:
SQL, Raw Data Discovery, ETL Transformación, Machine Learning.
BIG DATA
ETL
Pentaho: Carga Batch y Streaming. ETL y Orquestación.
Oozie: ETL y Orquestación.
Sqoop: Carga distribuida, Batch.
Kafka: Carga distribuida, Streaming.
FUENTES DE DATOS
ALMACENAMIENTO DISTRIBUIDO
Hadoop - HDFS
Amazon S3.
FUENTES DE DATOS
Bases de datos (RDBMS, Columnar, NoSQL), ficheros (json, xml, txt...), internet of the things (IoT), Apps, etc.

8. REFERENCIAS

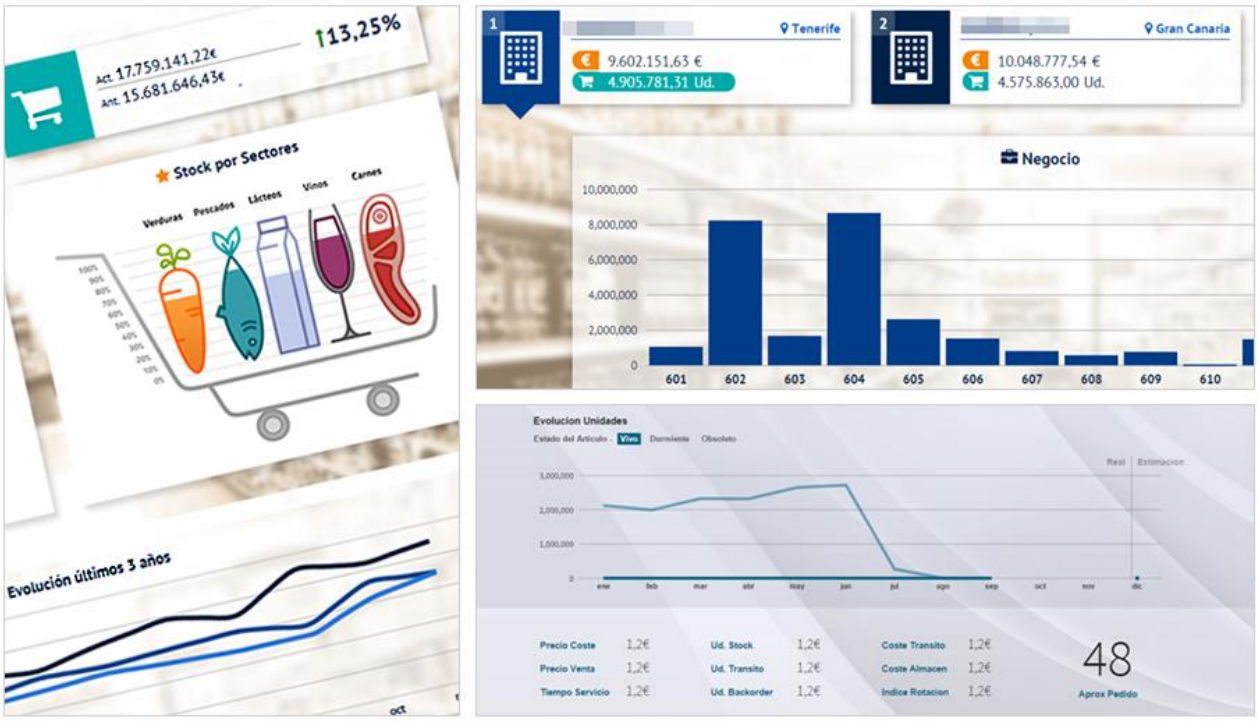
Trabajamos en los principales sectores y con algunas de las compañías y organizaciones más importantes de España.

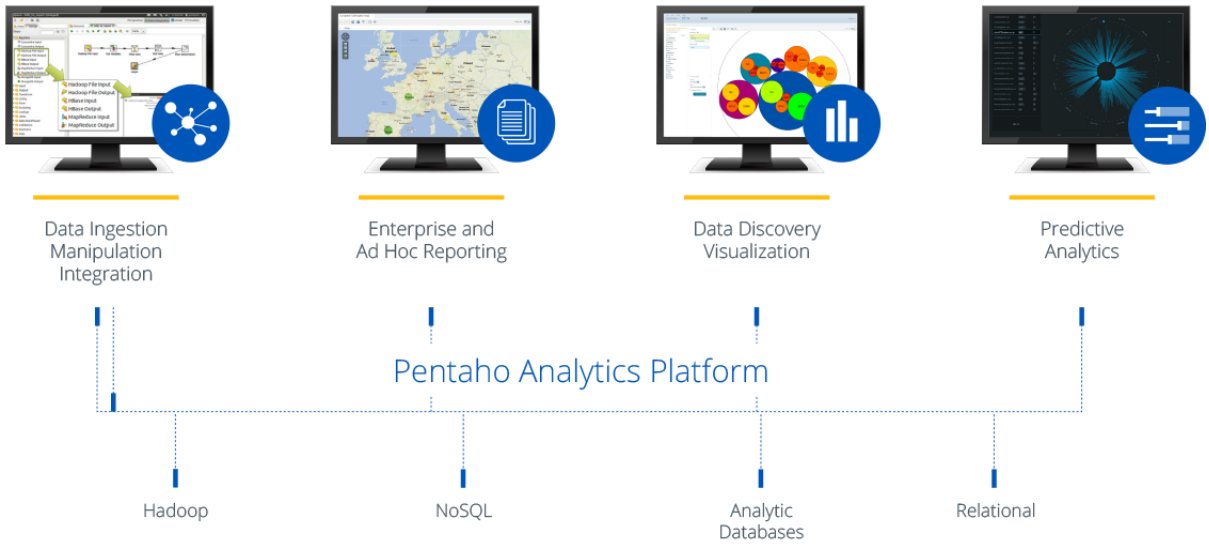
SECTOR PRIVADO

SECTOR PÚBLICO

9. EJEMPLOS DE DESARROLLOS ANALYTICS

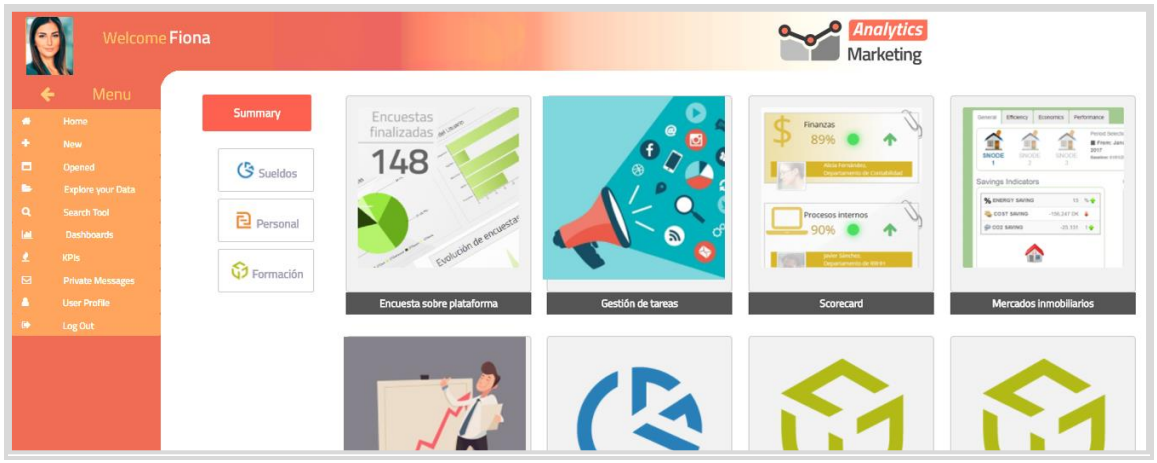
A continuación se presentan **ejemplos de algunos screenshots** de cuadros de mando diseñados por Stratebi, con el fin de dar a conocer lo que se puede llegar a obtener, así como Demos Online en la web de Stratebi:






facebook DASHBOARD COMENTARIOS
Sexo
Sentimiento
Localización
65%
35%
28%
60%
12%
Tipo de recurso
Evolución Comentarios por días
smart MADRID
Distrito | Barrio
Última Actualización
18:24
TOP 5 PUNTOS DE MEDIDA
PROXIMIDAD DE PUNTOS DE MEDIDA
EVOLUCIÓN INTENSIDAD DE TRÁFICO


Créditos Matriculados
Procedencia
FUENTE ÁLAMO
Alumnos Matriculados: 50
Sexo
50% Masculino
50% Femenino
Rango de Edad
18-20
20-22
22-24
24-26
26-28
28-30
30-35
35-40
GRADO EN BIOLOGÍA
Créditos Matriculados: 6.500
Buscar Plan de Estudio
INTERÉS PLATAFORMA
181 Inscritos en Cursos
8 Desinscritos en Cursos
181 Inicios de Cursos
110 Inscritos en Eventos Virtuales
63 Inscritos en Debates
181 Inscritos en Talleres
USUARIOS ACTIVOS POR PAÍS
INSCRITOS EN CURSOS POR COMUNIDAD EDU
JORNADA COMPLETA
IDIOMAS
TIC
TIPO DE CENTRO
TIPO DE AUSENCIA
Nº de profesores por sexo
36%
64%
Nº total de profesores
1.587
Media de edad
43 años
Búsqueda


Welcome Fiona
Analytics
Marketing
Menu
Home
New
Opened
Explore your Data
Search Tool
Dashboards
KPIs
Private Messages
User Profile
Log Out
Summary
Sueldos
Personal
Formación
Encuestas finalizadas
148
Evolución de encuestas
Finanzas
89%
Procesos internos
90%
Savings Indicators
Encuesta sobre plataforma
Gestión de tareas
Scorecard
Mercados inmobiliarios