Palenque
Plataforma de Grandes Datos para el Agro
Arquitectura de Big Data

Agustina Bazzano[1], Lautaro Chiarle[1],
Ernesto Mislej[2], Carlos Lizarralde[2], and Nicolás Higgs[2]

[1] Flux IT,
{agustina.bazzano, lautaro.chiarle}@fluxit.com.ar,
http://fluxit.com.ar/
[2] 7Puentes,
{ernesto, charly, nico}@7puentes.com,
http://7puentes.com/

Resumen Palenque es una plataforma y un ecosistema de aplicaciones que brindarán soluciones tecnológicas basadas en grandes datos a los productores agropecuarios, así como al sector público y otros actores del sistema productivo y científico. En este trabajo describiremos la arquitectura de la plataforma de datos.

Keywords: big data, agroTICs, plataforma de datos

Abstract. Palenque is a platform and an ecosystem of applications that provide technological solutions based on large data to farmers and the public sector, and other actors in the productive and scientific system. In this work, we will describe the architecture of the data platform.

Keywords: big data, agribusiness, big data platform

1 Sobre Palenque

En los últimos años ha comenzado a desarrollarse una nueva tendencia relacionada a la agricultura denominada *data-driven agriculture* en la cual la tecnología de la información se posiciona en el corazón de la explotación agrícola. Tecnologías vinculadas, por ejemplo, al sistema de geoposicionamiento satelital permiten obtener datos georeferenciales de distintos sitios de un lote, como ser imágenes satelitales, cartas de suelos, mapas topográficos, muestreo del suelo, rendimientos de cosechas anteriores, para luego poder procesar esta información con sistemas especializados, permitiendo elaborar diferentes mapas o modelos con información precisa en sus diferentes áreas. Las nuevas tecnologías permiten

realizar el seguimiento y la gestión de todas estas tareas de manera integrada, remota y en tiempo real, pudiendo regular sus acciones a partir del retorno que reciben.

Argentina, dada su reconocida competitividad en la agricultura, tiene en este sentido una significativa oportunidad. Es aquí donde aparece el Proyecto Palenque[1], el cual se lleva a adelante con una cooperación entre la Fundación Sadosky[2] y otros organismos públicos y privados.

2 Desafíos y requerimientos

La siguiente es una lista de los requerimientos funcionales que impactan sobre el diseño de la arquitectura de la plataforma de datos:

- Contar con un soporte para la ingesta y almacenamiento de datos provenientes tanto de instituciones (*mainstream*), con un flujo de carga controlado y normalizado; como de la comunidad de desarrolladores/productores (*long tail*).
- Brindar un sistema de control de acceso sobre datos privados.
- Favorecer la integración, la interacción y la colaboración de los distintos actores involucrados.
- Permitir el almacenamiento de datasets de gran volumen, series temporales de datos agrícolas, georreferenciados o no y documentación no-estructurada. Dar soporte para datos del tipo **Raster**.
- Permitir la integración, transformación y administración de los datos.
- Dar soporte para datos del tipo **Stream**.
- Proveer mecanismos para permitir el acceso a la plataforma datos pensada desde una perspectiva para el desarrollo de aplicaciones.
- Brindar servicios de procesamiento sobre los datos.
- Priorizar tecnologías abiertas.

3 Propuesta arquitectónica

Consideramos que el proyecto debe representar una referencia a nivel de arquitectura en cuanto a la implementación de soluciones abiertas. Es de suma importancia para el éxito que la plataforma provea mecanismos que faciliten el acceso a los datos y es ahí donde nos valemos de referencias como OpenAPI[3]. También consideramos importante que la plataforma provea los mecanismos de acceso a los recursos de la plataforma, a nivel de almacenamiento, procesamiento, visualización, cómputo y seguridad.

Como respuesta a los requerimientos planteados, se propuso una arquitectura basada en las siguientes componentes:

3.1 Cluster Hadoop

Hadoop[4] es el estándar técnico de facto para el desarrollo de plataformas de Big Data. Cuando hablamos de Hadoop no sólo hablamos de los servicios de almacenamiento y procesamiento en cluster, sino que hablamos del ecosistema de tecnologías que se enfocan en diferentes particularidades de la solución.

HDFS Es el *filesystem* distribuido de Hadoop. Provee la distribución y replicación de los files a través del cluster.

HBase Es la base de datos orientada a *very large tables*. Se monta sobre HDFS. La usaremos para brindar los servicios de acceso a los datos (planos no-estructurados), los datasets, series de datos temporales, etc.

Accumulo + Geowave Esta combinación de tecnologías permite la gestión de los datos del tipo raster.

Geoserver Es el clásico servidor de datos espaciales, esta vez montado sobre Accumulo + Geowave para dar soporte en Big Data.

Spark + YARN Spark como implementación del paradigma de programación distribuida sobre el cluster Hadoop. Utilizando como soporte de los datos tanto a HDFS como a YARN para la utilización de datos en memoria.

El siguiente gráfico representa un vista conceptual de alto nivel que contiene los componentes principales y su relación de integración y exposición:

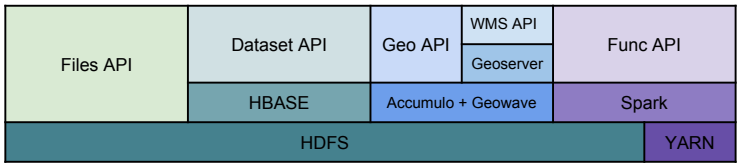

Figura 1. Esquema de componentes de arquitectura de la plataforma de datos.

3.2 API de servicios

Cada tecnología de base está encapsulada en una API Java, para simplificar el acceso a los servicios y facilitar su uso.

Files API Se encarga de exponer los servicios de lecto-escritura sobre el filesystem para datos archivos planos no-estructurados.

Dataset API Se utiliza para realizar consultas relacionales sobre los datasets, una estructura de datos diseñada para las consultas tempo-espaciales.

Geo API Se utiliza para acceder a datos geográficos, soportando operaciones de consulta del tipo geográficas.

WMS API La WMS API replica el protocolo clásico WMS para acceso a capas geográficas.

Func API Se utiliza para publicar y lanzar jobs del tipo Spark[5] sobre la plataforma de datos.

3.3 Imágenes Satelitales & Big Data

Estamos transitando una verdadera revolución en la industria satelital orientada a la provisión de imágenes y sensado remoto; apenas hace unos pocos meses, tanto la agencia USGS como la ESA liberaron sus respectivas imágenes Landsat[6] y Sentinel[7]. Respecto de la tecnología de nanosatélites, se encuentran en órbita 243[1] y durante el transcurso del año 2016, se prevén más de 400 lanzamientos, 2 de la argentina Satellogic[8]. La empresa Planet Lab[9] -que recientemente adquirió la alemana BlackBridge proveedora de RapidEye- prevé para el segundo semestre del 2016, proveer servicios de imágenes con un tiempo de revisita menor a los 3 días de revisita para toda la superficie terrestre del planeta. Satellogic tendrá revisita horaria en los próximos años cuando alcance su constelación de 300 nanosatélites. Teniendo en cuenta sólo la provincia de Buenos Aires, se registraron 696 escenas Landsat8 durante el año 2015; cada escena es de ∼1GB totalizando ∼700GB en 1 sólo año.

La arquitectura propuesta, para satisfacer los requerimientos GIS se basa en la librería Geowave[10]. Geowave provee los servicios de almacenamiento, indexación, búsqueda sobre datos multidimensionales -tanto desde el punto de vista espacial como temporal. Utiliza como soporte una DB key-value, siendo Accumulo[11] la DB de referencia[2]. Dentro de las características que brinda Geowave, destacamos:

Soporte de indexación multidimensional Lo utilizamos para modelar las serie de datos asignadas a coordenadas espaciales, como datos meteorológicos.

Soporte de objetos geográficos y operadores geoespaciales GeoWave intenta ser para Accumulo lo que PostGIS es a PostgreSQL.

Integración con GeoServer Implementa el protocolo de servicios OGC para visualizar y compartir datos.

Map-Reduce Implementa las operaciones de transformación de datos geográficos utilizando primitivas Map-Reduce para distribuir el procesamiento.

3.4 Seguridad de la Información

El control de acceso sobre los datos planos, datasets, rasters y demás recursos están alcanzados por una componente transversal de seguridad. Esta componente hoy se resuelve en una capa de middleware. Queda para una etapa futura, migrar la componente de seguridad al cluster.

4 Trabajo Futuro

Aún no conocemos cuál serán los casos de uso que requieran uso intensivo de los recursos, si la comunidad de usuarios se centrará más en el consumo de datos

[1] Al 18 de mayo de 2016

[2] En una futura versión de GeoWave soportará el uso de HBase

near-realtime o en los históricos, sobre cuáles regiones del país, etc. Debemos esperar para poder optimizar en esa dirección.

Quedará para versiones futuras trabajar con los requerimientos de indexación y trabajo con datos del tipo **Stream**. Se propone investigar tecnologías de indexación como SolR y de tratamiento de stream y colas como Kafka y Storm.

Asímismo fortalecer los aspectos de seguridad para extender su alcance sobre el cluster, un área poco resuelta en la comunidad.

Referencias

1. Fundación Sadosky: Proyecto Palenque. http://www.fundacionsadosky.org.ar/palenque/
2. Fundación Sadosky. http://www.fundacionsadosky.org.ar
3. The Open API Initiative (OAI). https://openapis.org/
4. The Apache Hadoop project. http://hadoop.apache.org/
5. Apache Spark. Lightning-fast cluster computing http://spark.apache.org/
6. USGS. Servicio Geológico de los Estados Unidos. Landsat 8, http://landsat.usgs.gov/landsat8.php
7. ESA. Agencia Espacial Europea. Misiones Sentinel, https://sentinel.esa.int/web/sentinel/home
8. Satellogic. http://www.satellogic.com/
9. Planet Labs. https://www.planet.com
10. GeoWave. https://ngageoint.github.io/geowave/
11. Apache Accumulo. https://accumulo.apache.org/