

Modelo Big Data

Modelo de Arquitectura Big Data - ICBF

El futuro es de todos
Gobierno de Colombia

BIENESTAR FAMILIAR

PÚBLICA

Agenda

- Conceptos Clave.
- Big Data.
  - Definición.
- Modelo de Big Data.
  - Objetivo del Modelo Big Data.
  - Modelo conceptual Big Data.
  - Modelo de Implementación Big Data.
  - Modelo TI Big Data.
  - Apalancamiento del Modelo Big Data en el marco del PETI y Plan de Transformación Digital.
  - Lineamientos de Aprovisionamiento Solución BigData.

BIENESTAR FAMILIAR

Conceptos Clave

Definiciones

- **DataSet**: Colecciones o grupos de datos relacionados que comparten el mismo conjunto de atributos.
- **Análisis de Datos**: Disciplina de examinar/explorar datos para encontrar hechos, relaciones, patrones y/o tendencias. Es realizado por personas.
- **Analítica**: Disciplina para entender datos analizándolos vía múltiples técnicas científicas y herramientas de automatización con el foco de encontrar patrones y correlaciones ocultas. Es realizado por programas de computación.
- **BI**: Proceso de obtener conocimiento para mejorar la toma de decisiones al interior de la empresa analizando datos externos y generados por sus procesos de negocio.
- **KPI**: Métrica que permite medir el cumplimiento del éxito dentro de un contexto particular.

BIENESTAR FAMILIAR

Tipos de Big Datos

- **Estructurados** -> Se ajustan a un modelo de datos o esquema.
- **No Estructurados** -> No se ajusta a un modelo o esquema, es inconsistente y no relacional.
- **Semi-Estructurados** -> Tiene un nivel de estructura y consistencia, no es relacional por naturaleza. La estructura se puede inferir de los propios datos.
- **MetaData** -> Contiene información acerca de las características y estructura de un dataset. No son propiamente un tipo de datos.

BIENESTAR FAMILIAR

Clasificación de Big Datos

- **Gold** -> Data con la máxima calidad, lista para su uso, por lo general obedece a un modelo canónico.
- **Silver** -> Data con cierto tratamiento y limpieza, calidad media y aceptable.
- **Bronce** -> Data con un bajo o nulo nivel de calidad y tratamiento, por lo general tan pronto se recibe de la fuente.

BIENESTAR FAMILIAR

Tipos de Analítica

- **Descriptiva**-> Responder preguntas alrededor de eventos que ya han ocurrido.
- **Diagnóstica** -> Se centra en determinar la causa de un fenómeno o acontecimiento que ocurrió en el pasado.
- **Predictiva** -> Trata de predecir resultados de eventos, basado en patrones, tendencias y excepciones encontrados en data histórica.
- **Prescriptiva** -> Basado en modelos predictivos para determinar acciones que deben ser tomadas, dada una situación determinada.

BIENESTAR FAMILIAR

Tipos de Análisis de Datos

- **Análisis Cuantitativo ->** Es una técnica que se centra en la cuantificación de los patrones y correlaciones encontradas en los datos. Basado en prácticas estadísticas.

- **Análisis Cualitativo** -> Se centra la descripción de diversas calidades de los datos utilizando palabras. Implica el análisis de una muestra pequeña en mayor profundidad en comparación con el análisis cuantitativo.

- **Minería de Datos** -> Se centra en el descubrimiento de datos en datasets masivos, utilizando técnicas basadas en software para identificar patrones y tendencias.

- **Machine Learning ->** Aprendizaje de máquina, que se centra en la generación de modelos predictivos.

BIENESTAR FAMILIAR

Big Data

BIENESTAR FAMILIAR

Big Data - Definición

Disciplina que se dedica a el análisis, procesamiento y almacenamiento de **grandes** colecciones de datos que se originan **frecuentemente** de fuentes **diversas**, utilizando metodologías y herramientas, para apoyar la toma de **decisiones**.

Soluciones de Big Data son requeridas cuando las técnicas y tecnologías tradicionales, para el procesamiento, almacenamiento y análisis de datos son insuficientes. Estás soluciones tiene un grado de madurez menor a las tradicionales e implican grandes niveles de artesanalidad.

BIENESTAR FAMILIAR

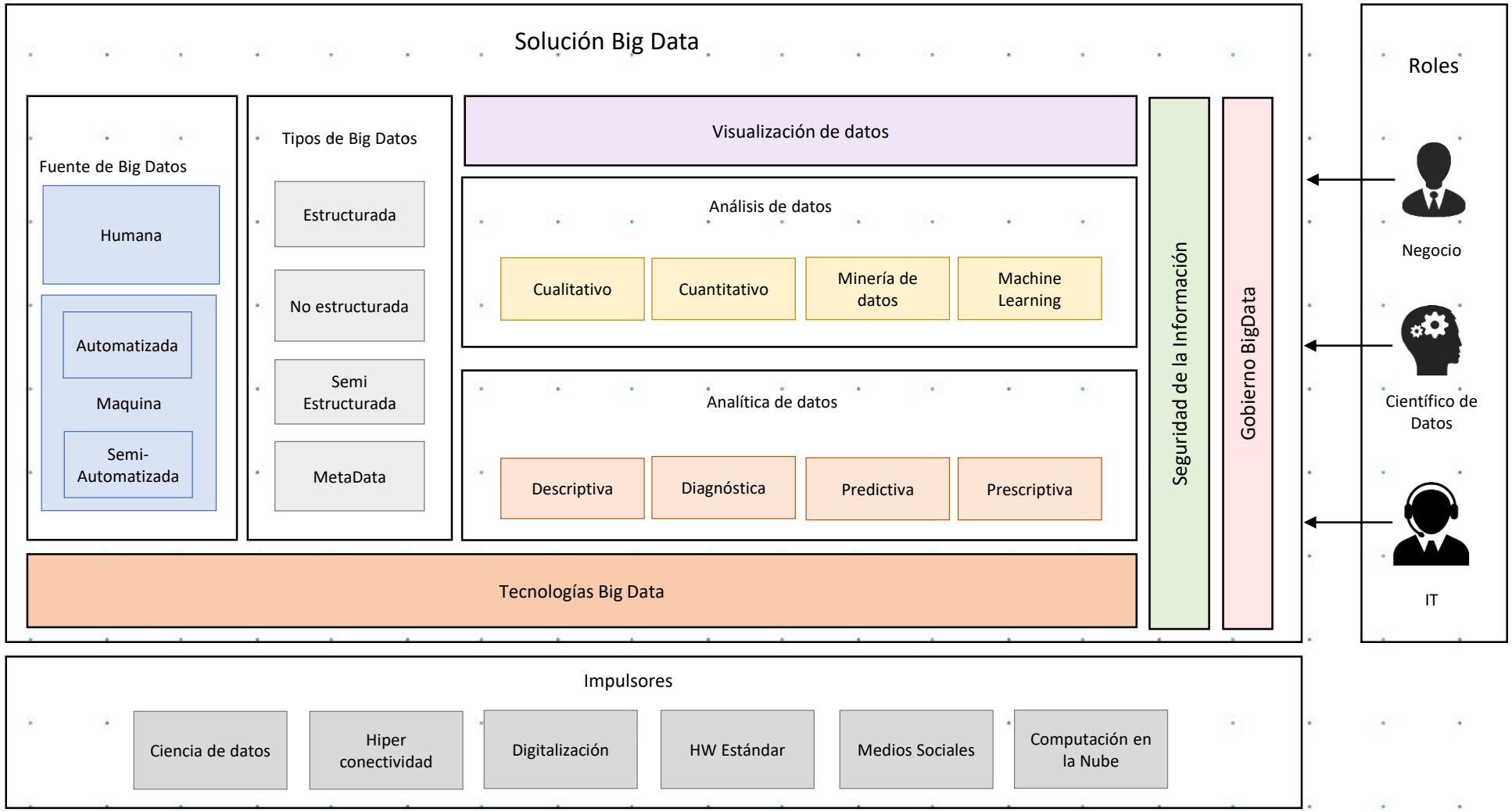
Modelo Big Data - ICBF

BIENESTAR FAMILIAR

Objetivo del Modelo Big Data ICBF

- ✓ Apalancar de manera estandarizada y estructurada tanto desde el punto de vista del negocio como técnico, cualquier iniciativa de Big Data que ese genere en el ICBF.

Modelo Conceptual Big Data

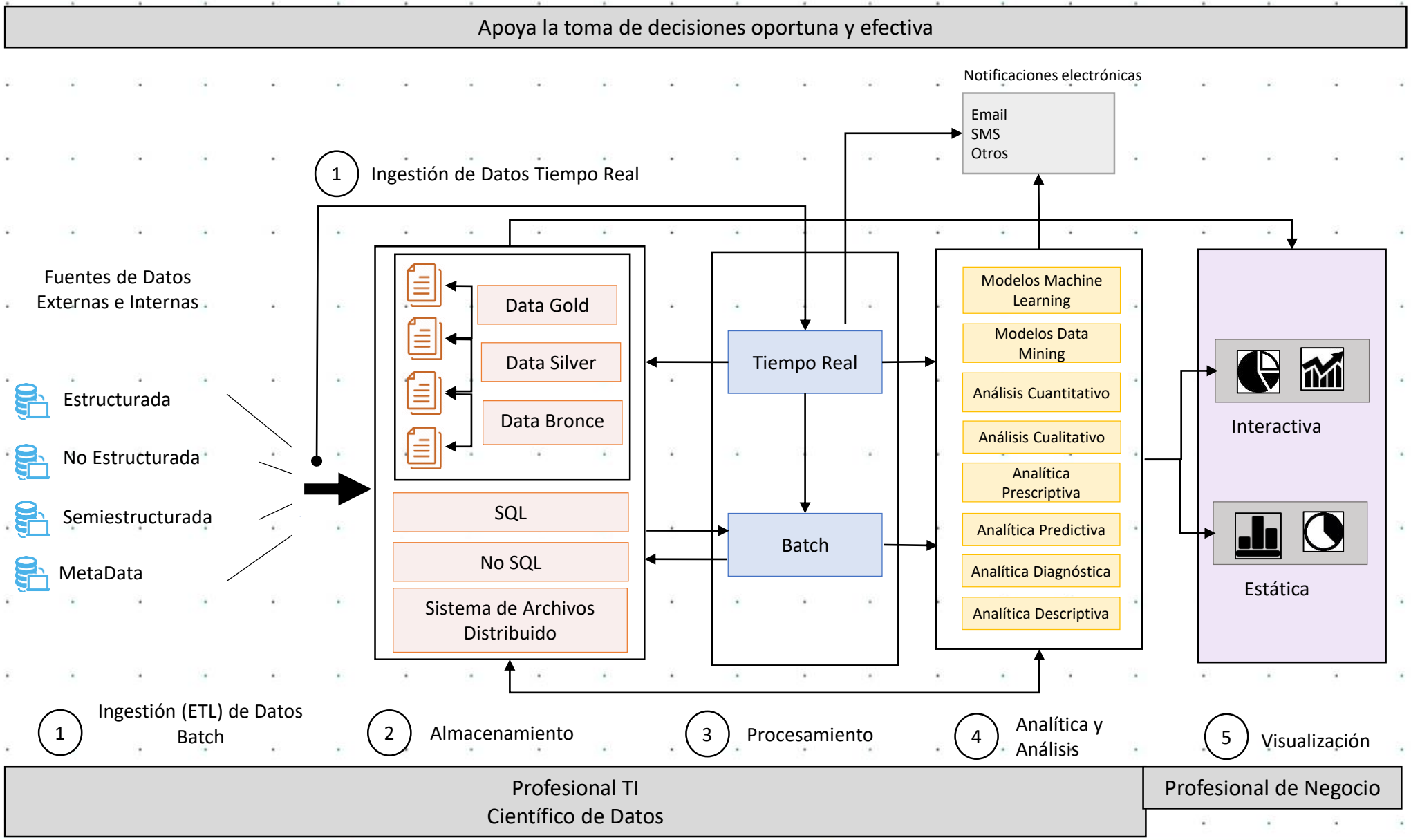


Modelo Conceptual Big Data

- Todo inicia con la generación de grandes cantidades de datos, clasificadas según su fuente de generación en humana (redes sociales, documentos de texto, contenido web, papers, artículos, etc) y máquina (IoT: sensores, logs, señales, etc), estas se almacenan y tipifican en Estructurada, No Estructurada, Semi-estructurada, MetaData. Con estos datos se realizan análisis (cualitativo, cuantitativo, minería de datos y machine learning) soportados en analíticas (descriptivas, diagnósticas, predictivas y prescriptivas) que apoyan la toma de decisiones del personal de negocio presentando los resultados y análisis en herramientas de visualización que permiten la interacción del usuario final. Todo lo anterior soportado por una plataforma tecnológica big data.
- Por último pero no menos importante, dos componentes transversales claves que complementan el modelo, todo el tema de seguridad de la información que debe ser tenido en cuenta y aplicado durante todo el flujo de ingesta, almacenamiento, procesamiento y visualización de los datos hasta convertirse en información que apoye la toma de decisiones oportuna y eficiente. El componente de gobierno que define, actualiza y evoluciones políticas, lineamientos, roles y responsabilidades que permitan que el modelo funcione según lo establecido.

BIENESTAR FAMILIAR

Modelo implementación Solución Big Data

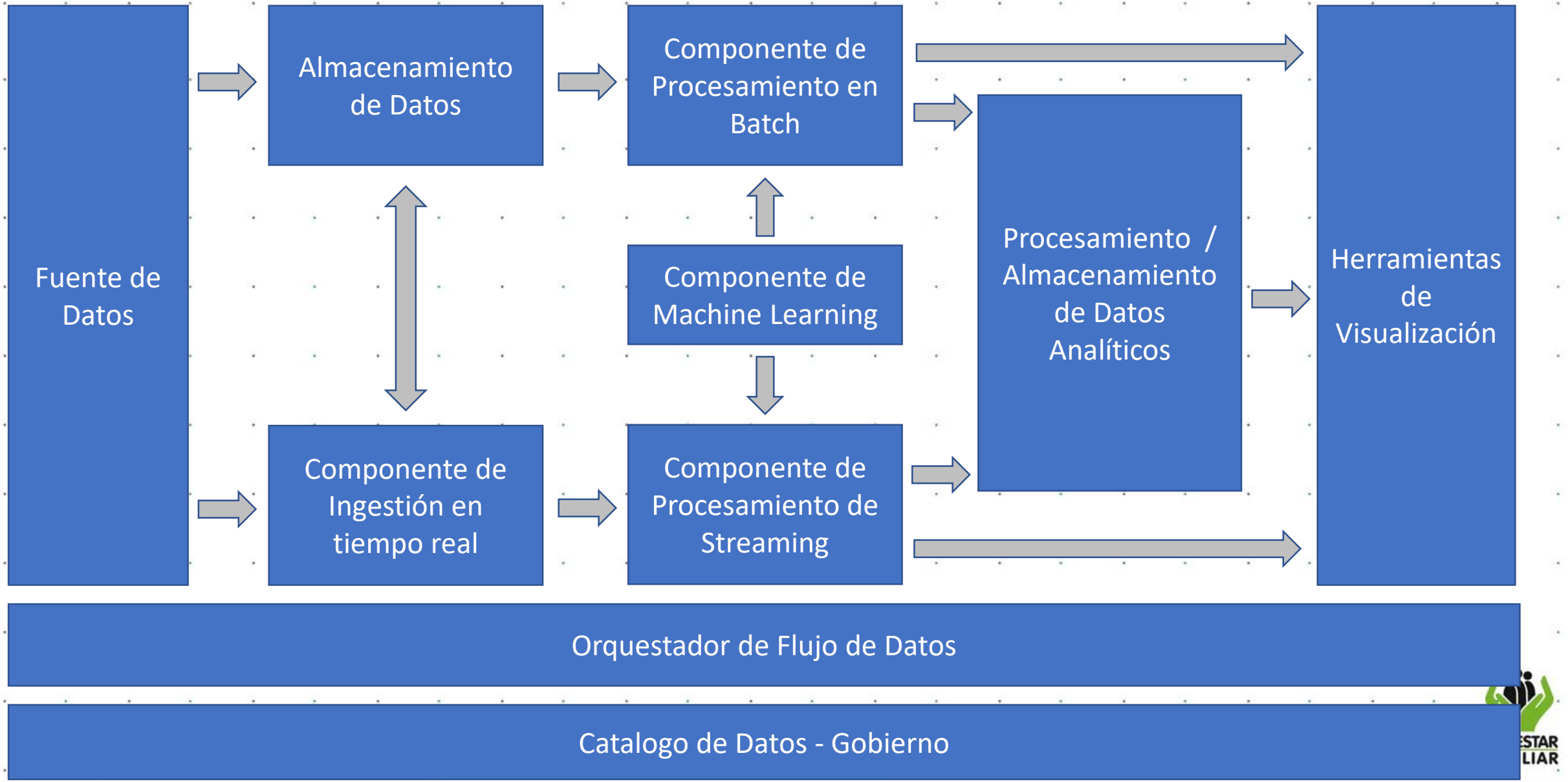


Modelo implementación Solución Big Data

1. La ingestión de datos, que es el proceso por el cual, se introducen los datos al modelo big data, se realiza de dos maneras, en batch (manera tradicional y más utilizada) o en tiempo real el cual soporta el procesamiento y analítica en línea.
2. El almacenamiento de grandes volúmenes, variedad y tipos de datos requiere de un modelo distribuido de archivos, bases de datos No SQL y/o base de datos SQL.
3. El siguiente paso es el procesamiento de la información que puede ser en batch o en línea, en este se realizan operaciones de extracción, agregación, validación y limpieza de datos para apoyar las labores de analítica y análisis de datos.
4. El análisis de datos se apoya en herramientas y técnicas analíticas (estadísticas, minería de datos y machine learning) sobre aquellos datos que han sido procesados previamente, adicionalmente sobre visualizaciones, donde también existen técnicas de visualización que permiten presentar la información de una mejor manera dependiendo de la naturaleza de los datos y lo que se quiere transmitir, que permitan a los analistas y científicos de datos realizar más eficientemente sus análisis.
5. Por último se encuentra la visualización de los datos, que es el paso donde se maximiza el valor real y tangible de la información de cara a los profesionales de negocio que la usan para apoyar la toma oportuna de decisiones.

BIENESTAR FAMILIAR

Modelo TI de la solución Big Data

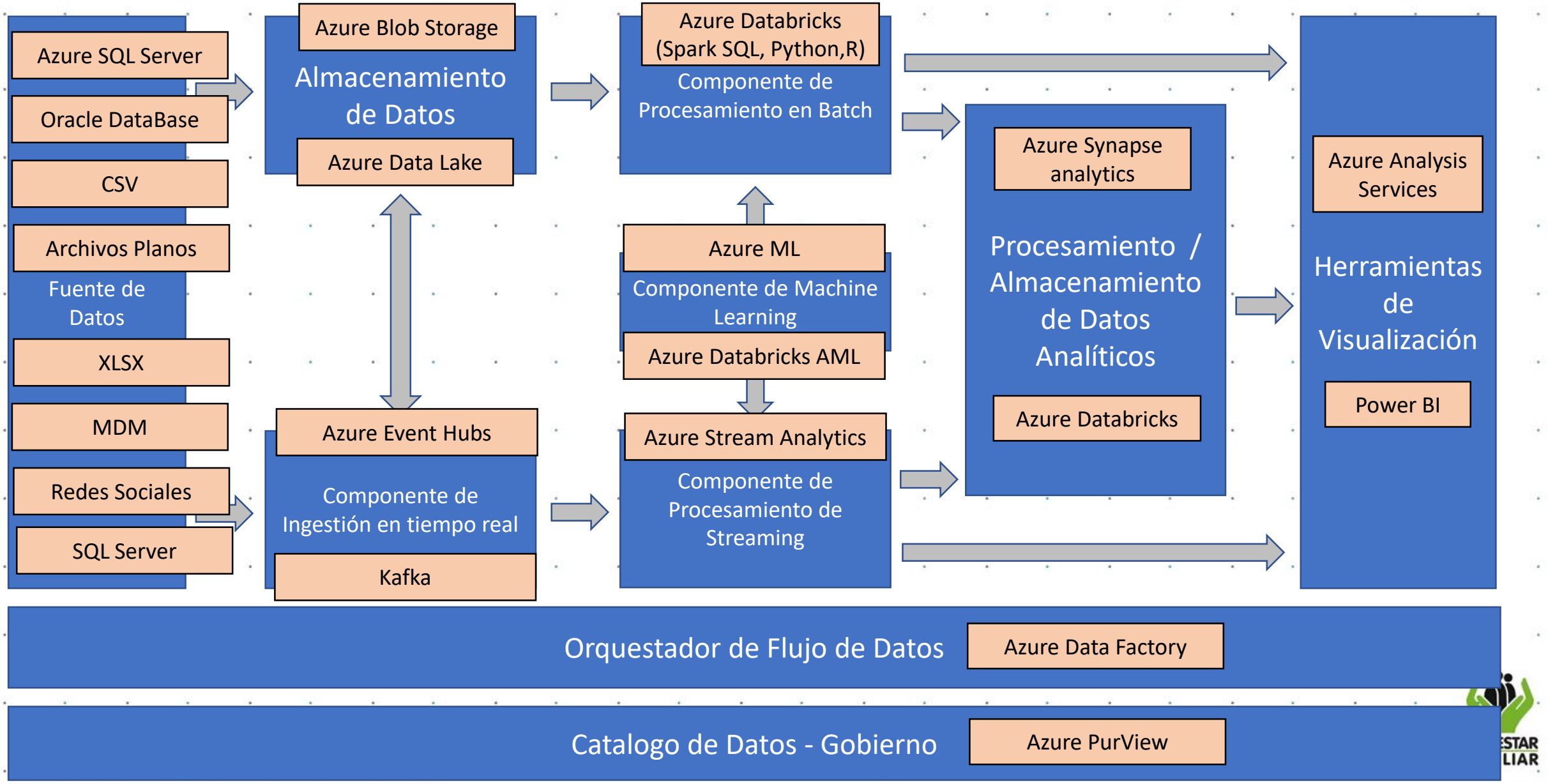


Modelo TI de la Solución Big Data

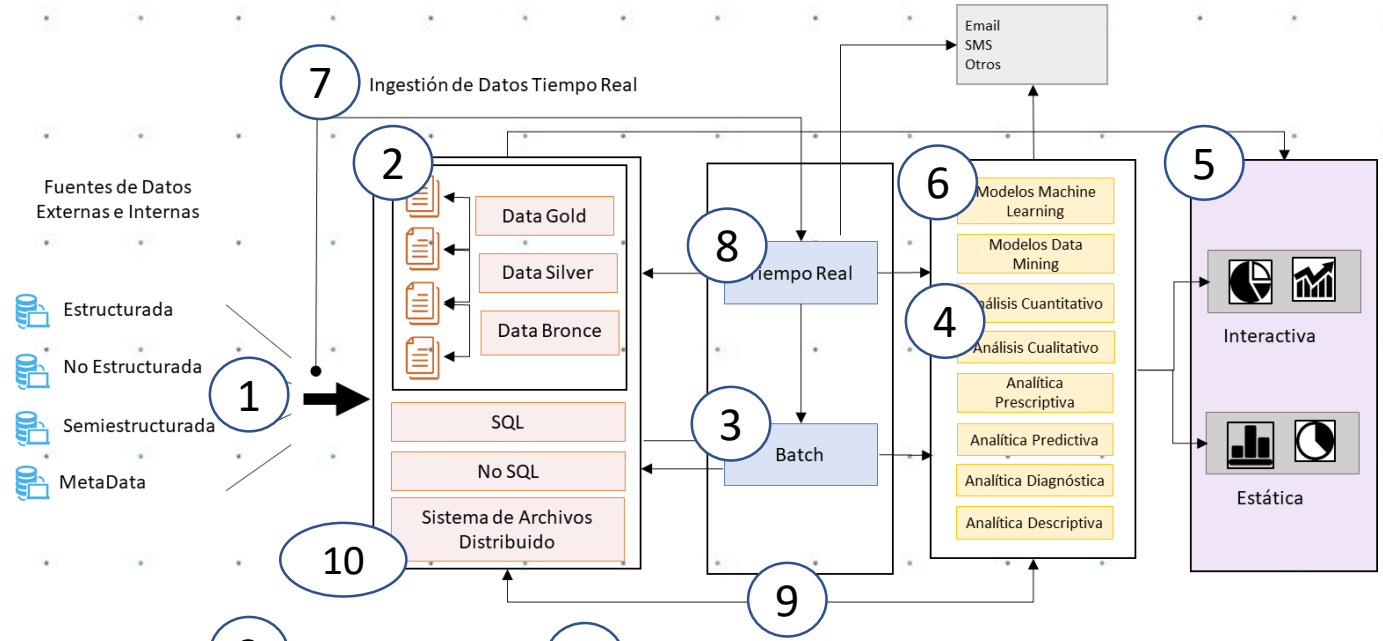
1. Almacenamiento de Datos: Se encarga de persistir los data sets con los diferentes tipos de datos, en bases de datos relacionales, no sql y/o sistemas de archivos distribuidos.
2. Componente de Ingestión en tiempo real: Es el encargado de procesar la data de los dispositivos que la generan en tiempo real (IoT devices).
3. Componente de Procesamiento en Batch: Es el encargado de procesar la data proveniente del sistema de almacenamiento (batch). Para el procesamiento en batch la regla es: Persisto primero, proceso después; para el procesamiento en tiempo real la regla es proceso primero, persisto después.
4. Componente de Procesamiento de Streaming: Es el encargado de procesar la data una vez se está ingestando en tiempo real.
5. Componente Machine Learning: Permite realizar analíticos de varios tipos de machine learning.
6. Procesamiento / Almacenamiento de Datos Analíticos: Permite realizar analíticos de varios tipos como son estadísticos, procesamiento de texto, procesamiento de lenguaje natural, además direcciona el almacenamiento de datos ya procesados y los deja disponibles para ser visualizados.
7. Orquestador de Flujo de Datos: Permite la coordinación de procesos (flujos de tareas) manteniendo la consistencia operacional. Es el componente encargado de ingresar y egresar los datos desde y hacia la solución tecnológica de big data.
8. Herramientas de Visualización: Permiten presentar de manera gráfica los resultados de los análisis realizados, adicionalmente soportan la interacción de los usuarios con la representación gráficas de los datos. Soportan la presentación de resultados como mapas de calor, series de tiempo, redes de trabajo, datos espaciales entre otras.
9. Catalogo de Datos: Permite catalogar los datos que se almacenan y que se producen durante los flujos de procesamiento de una solución BigData, esto con fines de búsqueda y organización de los datos.

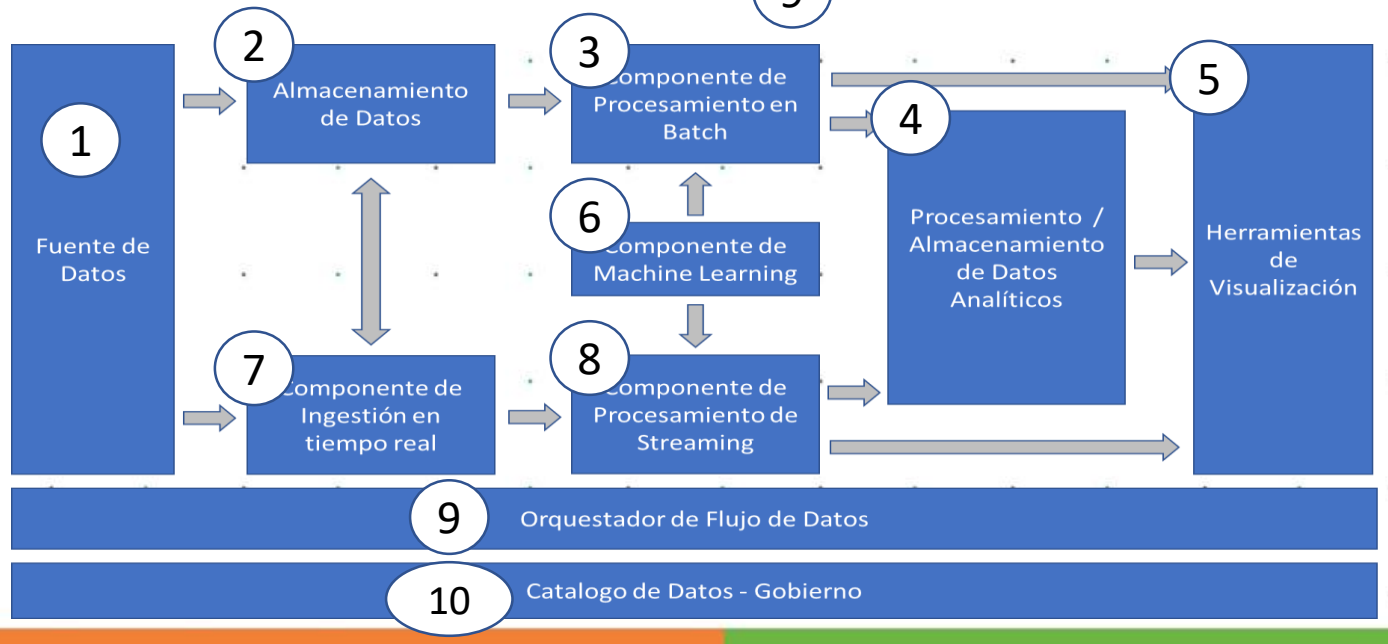
BIENESTAR FAMILIAR

Modelo TI de la solución Big Data vs Tecnologías



Modelo Implementacion Vs Modelo TI





1. MDM, SQL Server, Oracle, XLSX, CSV.

2. – Azure Blob Storage.
   – Azure Data Lake.
3. – Databricks (Spark SQL, Python,R).
4. – Synapse analytics
   – Azure Databricks.
5. – Azure Analysis Services
   – Power BI.
6. – Azure Machine Learning.
   – Azure Databricks AML.
7. – Azure Event Hubs.
   – Kafka.

8. – Azure Stream Analytics.

9. – Azure Data Factory.

10. – Azure Purview

BIENESTAR FAMILIAR

Lineamientos de Aprovisionamiento Solución BigData

1. - Blob Storage -> No requiere duplicar la Data, en pruebas se trabaja con la misma data que en producción.
   - Data Lake, Seguridad, organización.
      - Bronce -> Raw, files.
      - Plata -> DQ, data depurada.
      - Oro -> Semántico.

2. - Databricks (Spark SQL, Python,R)
   - Ambiente de Clusters por procesos (Calidad, Procesamiento de ML, Transformaciones).

3. - Databricks
   - Synapse Analytics, Azure SQL Server, DWH -> Ambiente independiente (Pruebas y Producción).

4. - Azure Analysis Services -> Ambientes independientes.
   - Power BI -> Workspace por ambiente o Deployment o APPs de Power BI.

5. - Machine Learning -> SaaS, PaaS -> Ambiente de procesamiento.

6. - Event Hubs -> Ambiente independiente (Storage, ADL) Pruebas / Producción.
   - Kafka

7. - Azure Stream Analytics -> Ambientes independientes leer data ingestada para pruebas.

8. - Azure Data Factory -> Pipelines por ambientes.

Grupos de Recursos Independiente por Proyecto y por ambiente.

BIENESTAR FAMILIAR



Gracias

El futuro es de todos
Gobierno de Colombia

BIENESTAR FAMILIAR

PÚBLICA