Método de orquestación de datos para servicios de fusión definidos por variables espacio-temporales.

José Carlos Morín García

Centro de Investigación de Estudios Avanzados
Unidad Tamaulipas

Dr. José Luis González Compeán
Dr. Iván López Arevalo

Presentación de Tesis, Agosto 2021

Contenido

Big Data

El tratamiento y análisis de enormes repositorios de datos, tan desproporcionadamente grandes que resulta imposible tratarlos con las herramientas de bases de datos y analíticas convencionales[1].

Big Data trata de tres cosas:

1. Las técnicas y la tecnología.
2. Escala extrema de datos que supera a la tecnología actual debido a su volumen, velocidad y variedad.
3. El valor económico, ayudando a la toma de decisiones.

Datos Meteorológicos

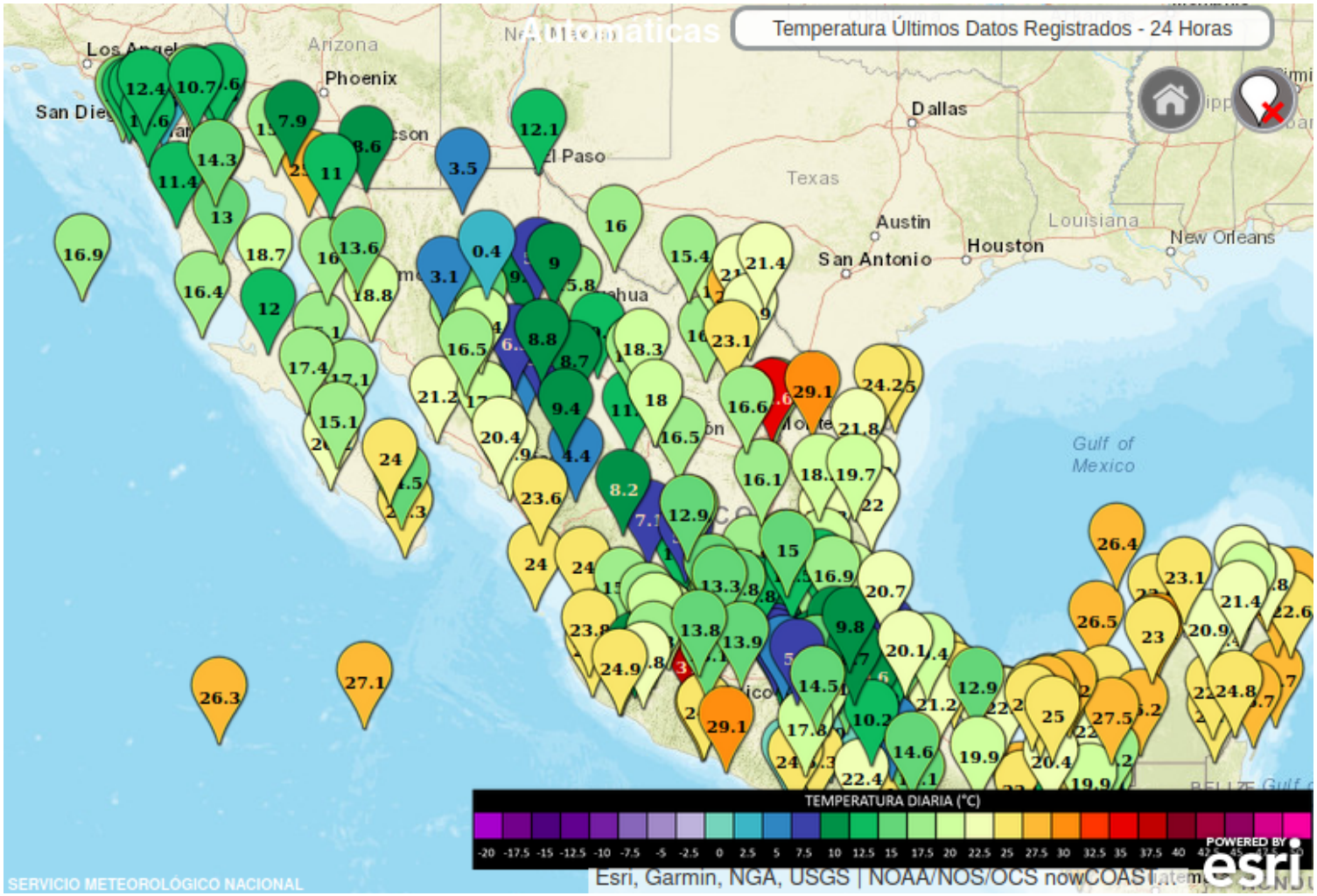

Figure: Antenas de EMAS [2].

Fusión de Datos (*FD*)

Definición

Se define como el conjunto de métodos, herramientas y medios que utilizan datos provenientes de diversas fuentes de forma heterogénea para unirlos en un solo conjunto de datos [3].

Objetivo

La fusión de datos tiene como objetivo obtener un resultado de mejor calidad, de múltiples fuentes, eventualmente heterogéneas, haciendo inferencias que pueden no ser posibles de una sola [4].

Orquestación de Datos

Es un proceso automatizado en el que una solución de software combina, limpia y organiza datos de múltiples fuentes, luego los dirige a servicios posteriores donde varios equipos internos pueden utilizarlos [5].

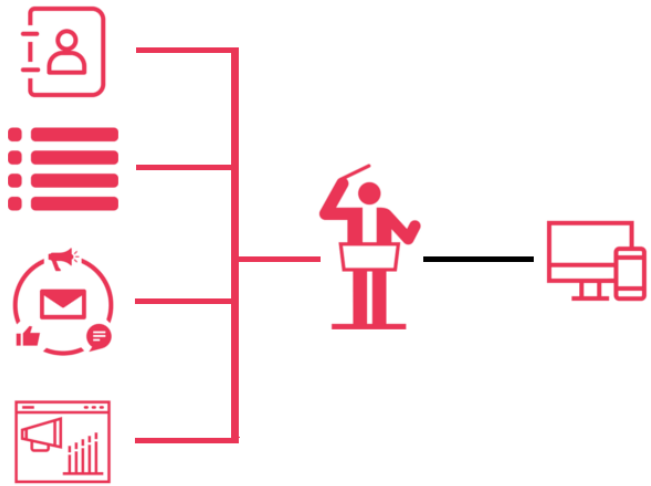

Figure: Proceso de Orquestación de Datos

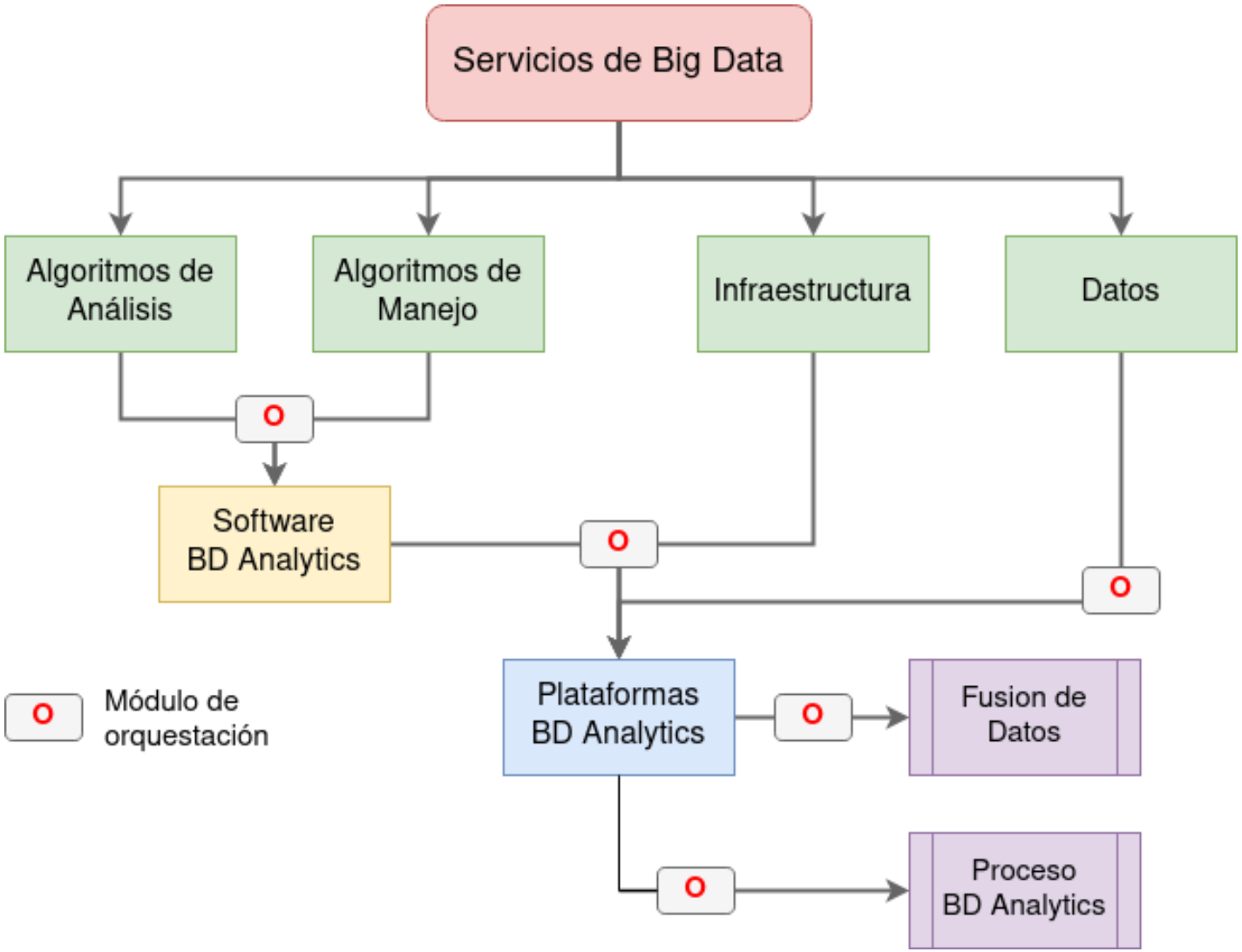

Figure: Servicios de Big Data

Problemática

Problemas

- *Dependencia Usuario-Proveedor*: El *vendor lock-in* se produce cuando el usuario delega tanto las fuentes de datos, como los servicios de *BDA* y orquestación a los proveedores de servicios en la nube.
- *Dependencia Fusión-Orquestación*: Se produce cuando el usuario solo crea procesos de *FD* dependiendo de los servicios de *BDA* que contiene el catálogo del proveedor y datos disponibles en su nube, los cuales están definidos por un esquema de orquestación propio y predefinido.

Problematica

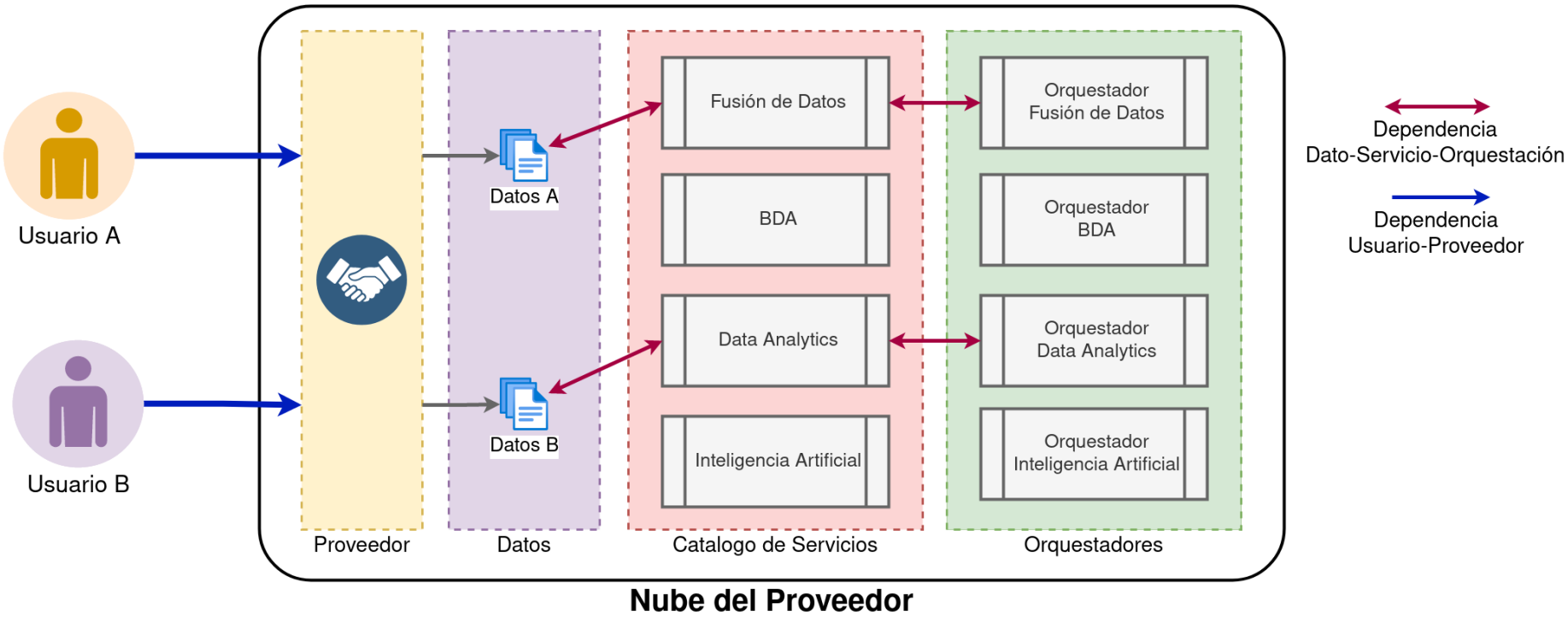

Figure: Dependencias entre el usuario y los servicios de Big Data

Preguntas de Investigación

Preguntas

- ¿Qué método de orquestación permitiría crear servicios de *FD* que mitiguen situaciones de *vendor lock-in*?
- ¿Qué etapas del método de *FD* permitiría disminuir la dependencia de Fusión-Orquestador?

Premisas

Dependencia Usuario-Proveedor

1. Los usuarios deben crear sus servicios *FD* asumiendo como límites los recursos ofrecidos por el proveedor.
2. El proveedor determina los factores no-funcionales (seguridad, confiabilidad, eficiencia, costos, etc).

Premisas Parte 2

Dependencia Dato-Servicio-Orquestador

1. Las variables espacio-temporales son comúnmente usadas para crear la *FD*.
2. La orquestación de datos está ligada funcionalmente (por las tareas que se realizan) a la infraestructura, plataforma y/o software empleados.
3. Los métodos de *FD* dependen funcionalmente de la orquestación definida por el proveedor.

Hipótesis

Hipótesis

Un método de orquestación, independiente de la infraestructura y determinado por esquemas basados en variables espacio-temporales, podría crear servicios de FD dinámicos que mitiguen los efectos de las dependencias Usuario-Proveedor y Fusión-Orquestador

Objetivos

General

Crear un método de orquestación para servicios de *FD* definidos por variables espacio-temporales.

Particulares

1. Definir un conjunto de cláusulas declarativas para definir un modelo de procesamiento agnóstico (de la infraestructura y/o plataforma) para crear servicios de *FD* sobre múltiples entornos de nube.
2. Crear un esquema de despliegue y acoplamiento de servicios de *FD* definido por variables espacio-temporales empleando el modelo anterior.

Cinvestav

Solución propuesta

Dependencia Usuario-Proveedor

Se creará un método de despliegue y acoplamiento de servicios *FD* que permitirá a los usuarios finales crear estos servicios mediante un esquema de cláusulas declarativas (por definir).

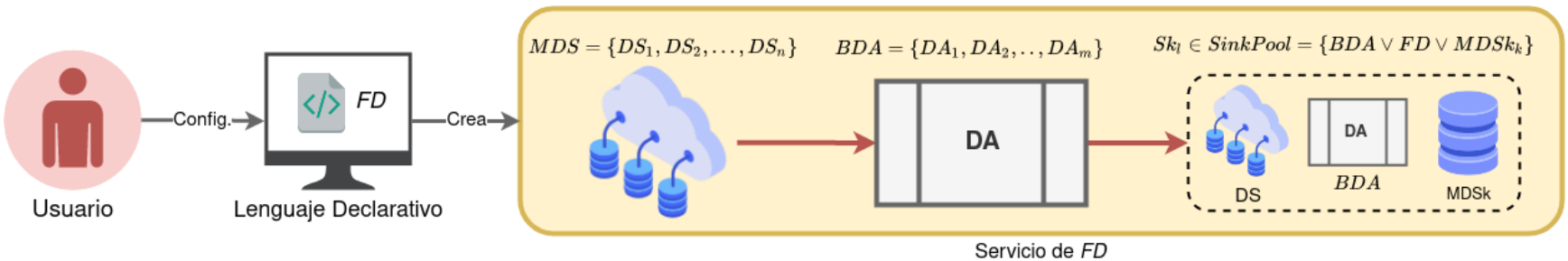

Figure: Solución Dependencia Usuario-Proveedor.

Solución Prouesta Cont.

$$FD = \{MDS_{i=1}^{n} \xrightarrow{D_i} BDA \xrightarrow{R_i} Sk_l\}$$

Donde:

$$MDS = \{DS_i, DS_{i+1}, \ldots, DS_n\}, (1 < n)$$
$$DS_i = (Id, Path), DS_i \neq \emptyset$$
$$BDA = \{DA_j, DA_{j+1}, \ldots, DA_m\}, (1 \leq m)$$
$$DA_j = (Id, Path) \wedge DA_j \neq \emptyset$$
$$MDSk = \{DSk_k, DSk_{k+1}, \ldots, DS_o\}, (1 \leq o)$$
$$DSk_k = (Id, Path) \wedge DSk_k \neq \emptyset$$
$$Sk_l \in SinkPool = \{BDA \vee FD \vee MDSk_k\}$$

Solución propuesta Cont.

El significado de las variables son las siguientes:

- MDS = Múltiples fuentes de datos (*Multiple Data Source*)
- DS = Fuente de datos (*Data Source*)
- BDA = Conjunto de procesos de Big Data Analytics (*Big Data Analytics*)
- DA = Análisis de Datos (*Data Analytics*)
- $MDSk$ = Múltiples depósitos de datos (*Multiple Data Sink*)
- DSk_k = Fuente para depósito de datos (*Data Sink*)
- $SinkPool$ = Conjunto de depósitos de datos (*SinkPool*)
- Sk_l = Tipo de depósito de datos

Solución propuesta Cont.

Dependencia Fusión-Orquestador

A partir del servicio de *FD* creado por el usuario con el esquema declarativo, crear un servicio de orquestación que pueda desplegar múltiples versiones del servicio de *FD* (por definir).

Solución propuesta Cont.

$$X = spatioPatt = \{altitudX, latitudY\}$$
$$Y = tempo = \{Dia = 1 \wedge Mes = 1\}$$
$$Z = (X_i, Y_i = DS_i)$$

$Eventos = \{Temp.\,Max, Temp.\,Min, Humedad\}$

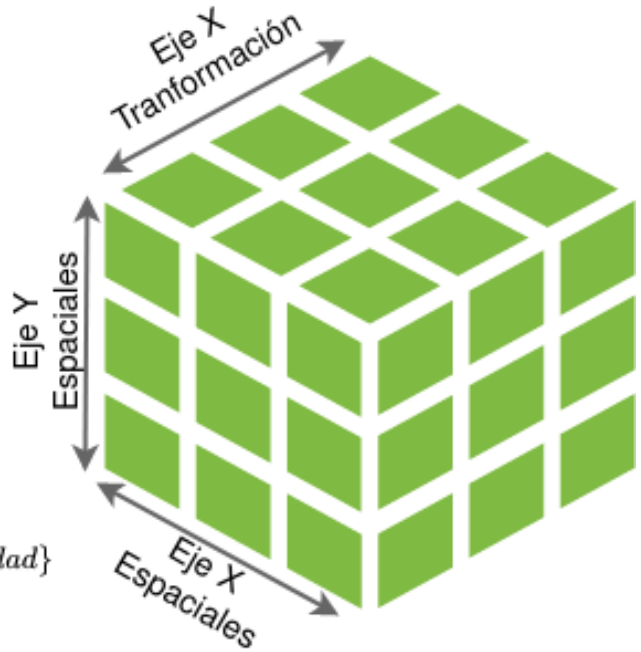

Figure: Cubo espacio-temporal.

Ejemplo de un servicio de *FD*

Esquema conceptual propuesto.

Metodología

- **Fase 1:** Diseño e implementación de la solución propuesta.
- **Fase 2:** Experimentación.
- **Fase 3:** Análisis de Resultados.
- **Fase 4:** Redacción de documentación.

Estado del Arte

En la Tabla 1 se ilustran las características identificadas como comunes entre estos trabajos:

1. *Enfoque de desarrollo*.
2. *Posibilidad de añadir nuevas fuentes de datos*.
3. *Reducen los escenarios de Vendor Lock-in*.
4. *Estandarización de datos personalizada*.
5. *Permite datos no estructurados*.
6. *Monitoreo*.

Estado del Arte Cont.

	Google Cloud	**AWS Glue**	**Airflow**	**Spatial Linked**	**Healthcare**	**Spatial *FD***	**Propuesta**
1	Integración de Datos	ETL, Integración de Datos	Orquestación, *Workflows*	Fusión de datos, Orquestación de *Workflows*	Fusión de datos, Orquestación de *Workflows*	Fusión de datos	Fusión de datos, Orquestación de Datos
2	Posible	Posible solo con códigos en Scala o Python	Posible	No Posible	No Posible	No Posible	Posible
3	Reduce con CDAP [**?**]	Medianamente Posible	No especificado	No Posible	No Posible	No Posible	Posible
4	Posible	Posible	Posible	No posible	No Posible	No Posible	Posible
5	Posible	Posible	Posible	No posible	No Posible	No Posible	Medianamente Posible
6	Posible	No, pero posible	Posible	Posible	No Posible	Posible	Posible

Table: Comparación entre los diferentes trabajos.

Contribuciones

Contribuciones esperadas

- Un nuevo método para la orquestación de datos basado en variables espacio-temporales configurables para crear servicios de fusión de datos independiente de la infraestructura y/o plataforma.
- Un modelo de procesamiento declarativo genérico para la construcción de servicios de análisis de datos basados en variables espacio-temporales.
- Un prototipo funcional que implemente el método propuesto.

¿Preguntas?

Gracias por su atención

Referencias I


Juan Jose Camargo-Vega, Jonathan Felipe Camargo-Ortega, and Luis Joyanes-Aguilar.
Conociendo Big Data.
Revista Facultad de IngenierÃa, 24:63 – 77, 01 2015.


Desarrollo.
Estaciones meteorológicas automáticas (emas).


Marc Mangolini.
Apport de la fusion d'images satellitaires multicapteurs au niveau pixel en télédétection et photo-interprétation.
1994.

Referencias II

Anna de Juan and R. Tauler.
Chapter 8 - data fusion by multivariate curve resolution.
In Marina Cocchi, editor, *Data Fusion Methodology and Applications*, volume 31 of *Data Handling in Science and Technology*, pages 205 – 233. Elsevier, 2019.

What is data orchestration amp; why it's essential for analysis.