

BIG DATA

David Martínez Martín

¿Qué es Big Data?



“**Big Data**” son aquellos datos cuya escala, diversidad y complejidad requieren de nuevas arquitecturas, técnicas, algoritmos y analíticas para extraer el **valor** y el **conocimiento** oculto en ellos.



Big Data - Definición



El **Big Data** o **Datos Masivos** es un concepto que hace referencia a la acumulación de grandes cantidades de datos y a los procedimientos usados para encontrar patrones repetitivos dentro de esos datos.

El fenómeno del **Big Data** también es llamado **datos a gran escala**.



Conjuntos de datos extremadamente grandes que pueden ser **analizados computacionalmente** para revelar patrones, **tendencias** y asociaciones, especialmente relacionadas con el **comportamiento** humano y las **interacciones**.

Big Data – Casos de Uso



Amazon almacena tus búsquedas y las correlaciona con las búsquedas de otros usuarios para generar recomendaciones personalizadas a tus gustos y necesidades.



Big Data – Casos de Uso



Walmart, en 2004 antes del paso del huracán Frances por las costas de Florida, realizó el análisis del histórico de ventas durante los días previos a emergencias naturales pasadas permitió determinar qué productos incrementan su demanda durante esos días.

Además de productos básicos como agua embotellada, el modelo logró predecir un aumento de hasta siete veces en la tasa de venta de los **pop tarts de fresa**, y que el producto más vendido sería la **cerveza**.



Big Data – Casos de Uso



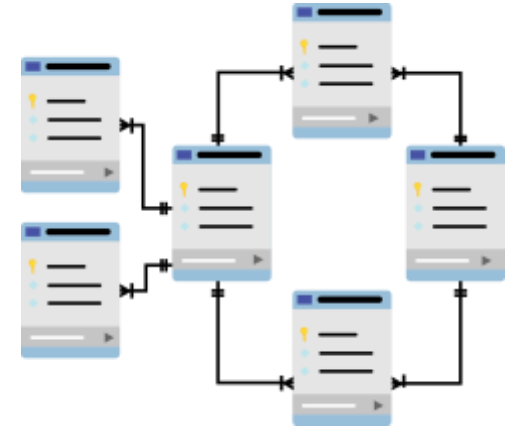
Netflix analiza los patrones de visionado de películas y series de sus usuarios para entender sus gustos e intereses. Esto les permite decidir qué series originales producir.



Big Data – Desafíos

Big Data implica grandes desafíos:

- El tamaño del “Big Data” supera la **capacidad de almacenamiento** y procesamiento de las bases de datos relacionales
- Gran porcentaje de los **datos no son estructurados**
- Análisis en tiempo real
- Visualización de grandes cantidades de datos

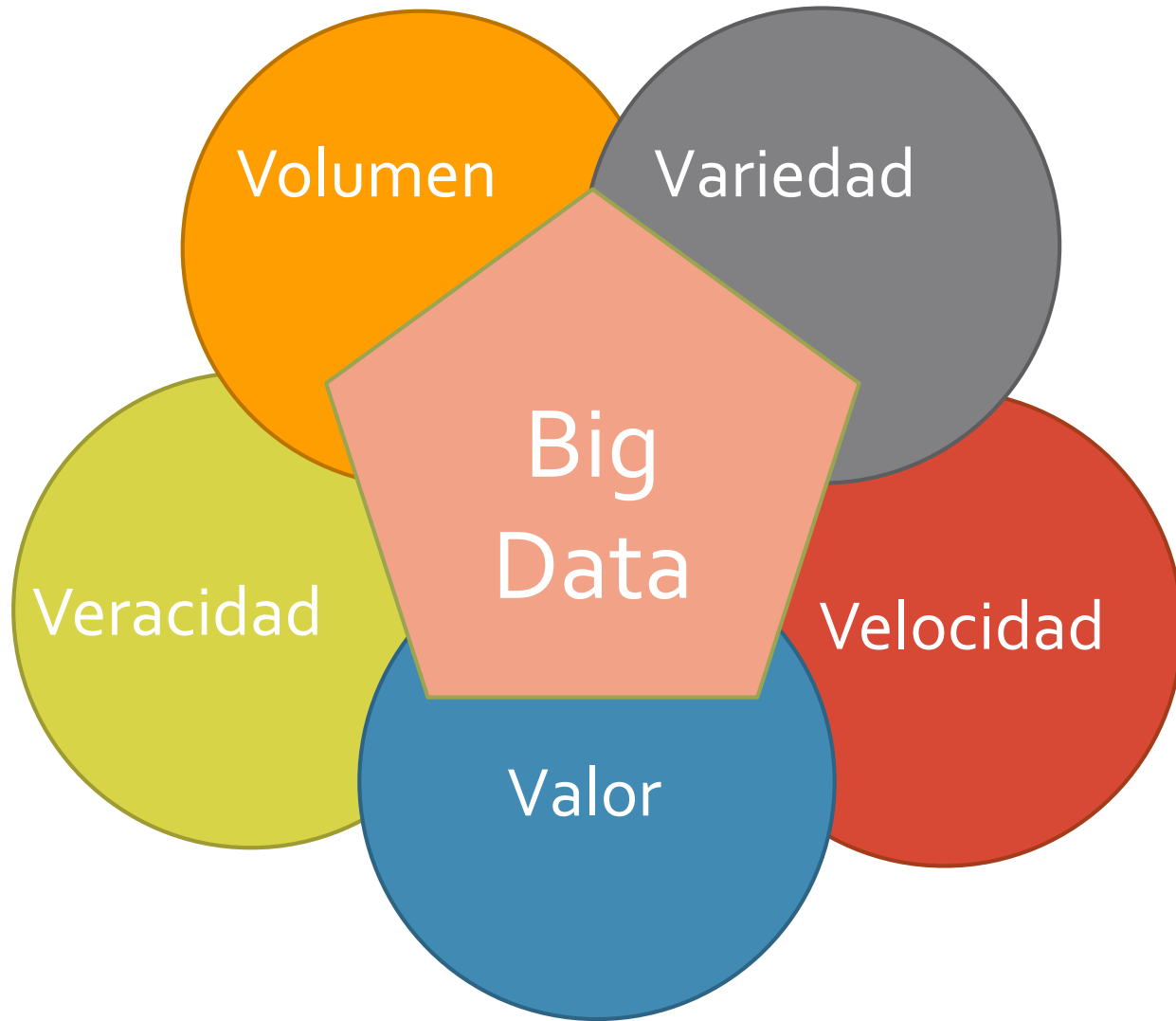


Big Data vs Business Intelligence

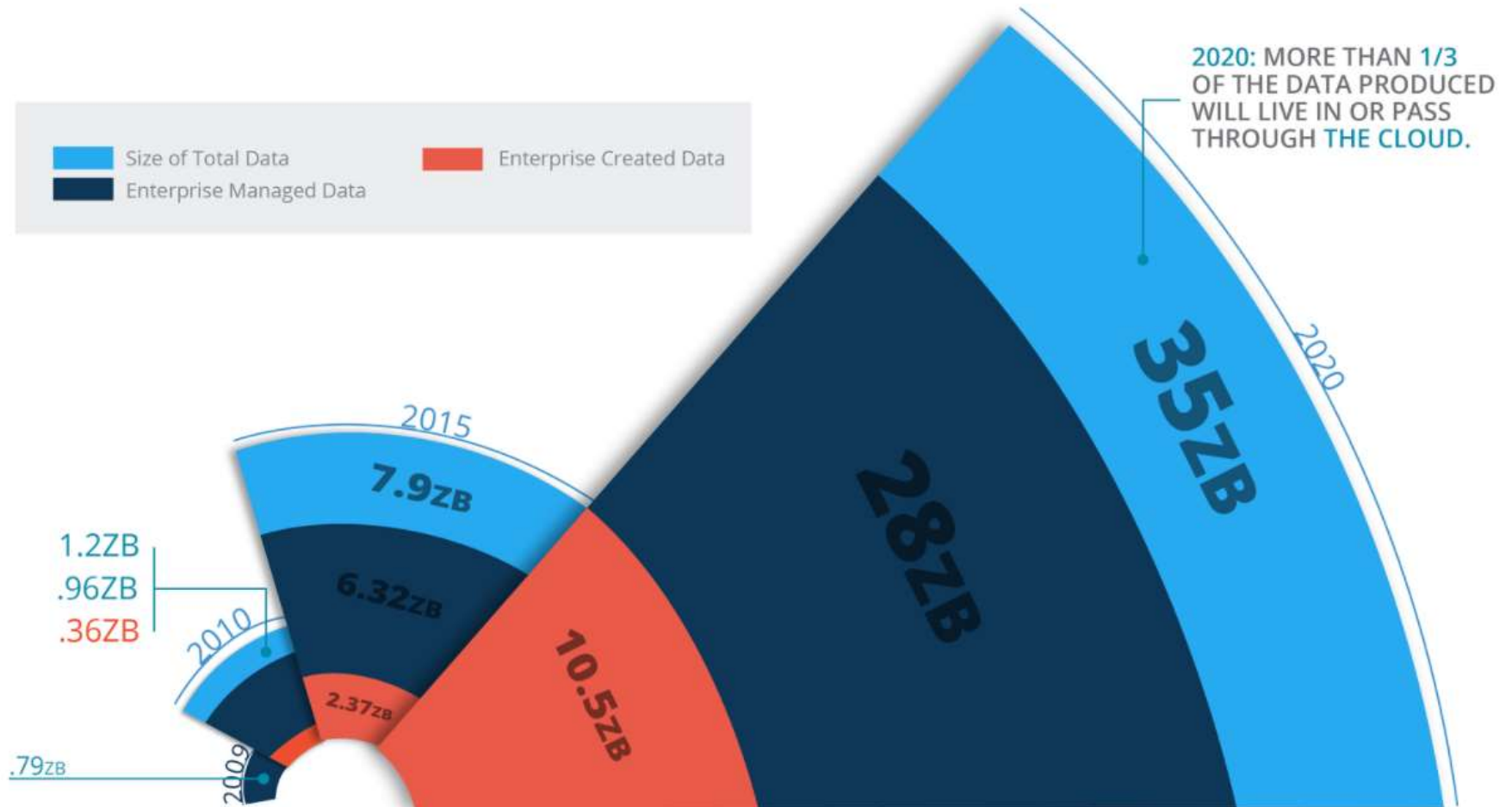
Business Intelligence	Big Data
Almacenamiento en un servidor central	Almacenamiento distribuido
Datos → Funciones de procesamiento	Funciones de procesamiento → Datos
Datos estructurados	Datos estructurados y no estructurados
Datos históricos	Datos históricos y en tiempo real
Procesamiento secuencial	Procesamiento paralelo
Escalado vertical	Escalado horizontal



Big Data – 5V's



Big Data – 5V's: Volumen



Big Data – 5V's: Variedad

Structured

- 20% de los datos
- Finanzas, ventas, almacén

Semi-Structured

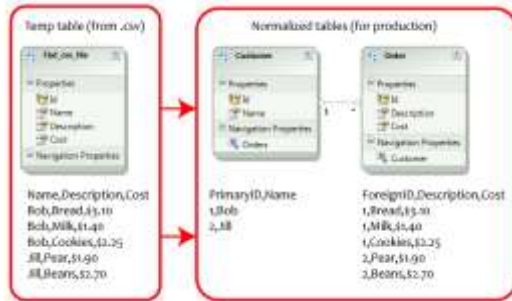
- JSON, XML (Parser)
- MongoDB, Cassandra...

Unstructured

- 80% de los datos
- Imágenes, audio, redes sociales..

Big Data – 5V's: Variedad

Estructurados



No estructurados



Data Lake



Big Data – 5V's: Velocidad

Los datos se generan muy rápido y necesitan ser procesados a una gran velocidad.

- **Procesamiento Batch:** Los datos se acumulan y procesan periódicamente.
- **Procesamiento Streaming:** Los datos se procesan de forma inmediata y requieren una arquitectura de baja latencia.



Big Data – 5V's: Valor

Las empresas pueden explotar los datos de sus clientes para obtener beneficios y generar **Valor** para:

- Mejorar las estimaciones del riesgo de crédito
- Aumentar la cartera de clientes
- Aminorar la pérdida de clientes
- Aumentar la satisfacción de los clientes
- Incrementar la eficiencia de los programas de marketing
- Aumentar las ventas mediante análisis predictivos



Big Data – 5V's: Veracidad

Estandarización:

Si no se unifican, en los análisis se considerarían poblaciones distintas



Población	Población estandarizada
MADRID	Madrid
MADRD	Madrid
Madri	Madrid
MAadrid	Madrid
Mmadrid	Madrid
MAD	Madrid
Madrid Centro	Madrid
Madrid	Madrid

Big Data – 5V's: Veracidad

Datos duplicados

ID	Nombre	Fecha de nacimiento	Domicilio	Población	Código Postal
23	Pedro Gómez	10/12/1965	Avenida España, 36	TERRASSA	8224
67	Sr. Pedro Gómez	10/12/1965	AV ESPAÑA	Terrassa	8224
95	PEDRO GOMEZ	10/12/1965	Avenida España, 36	Terrassa	8224



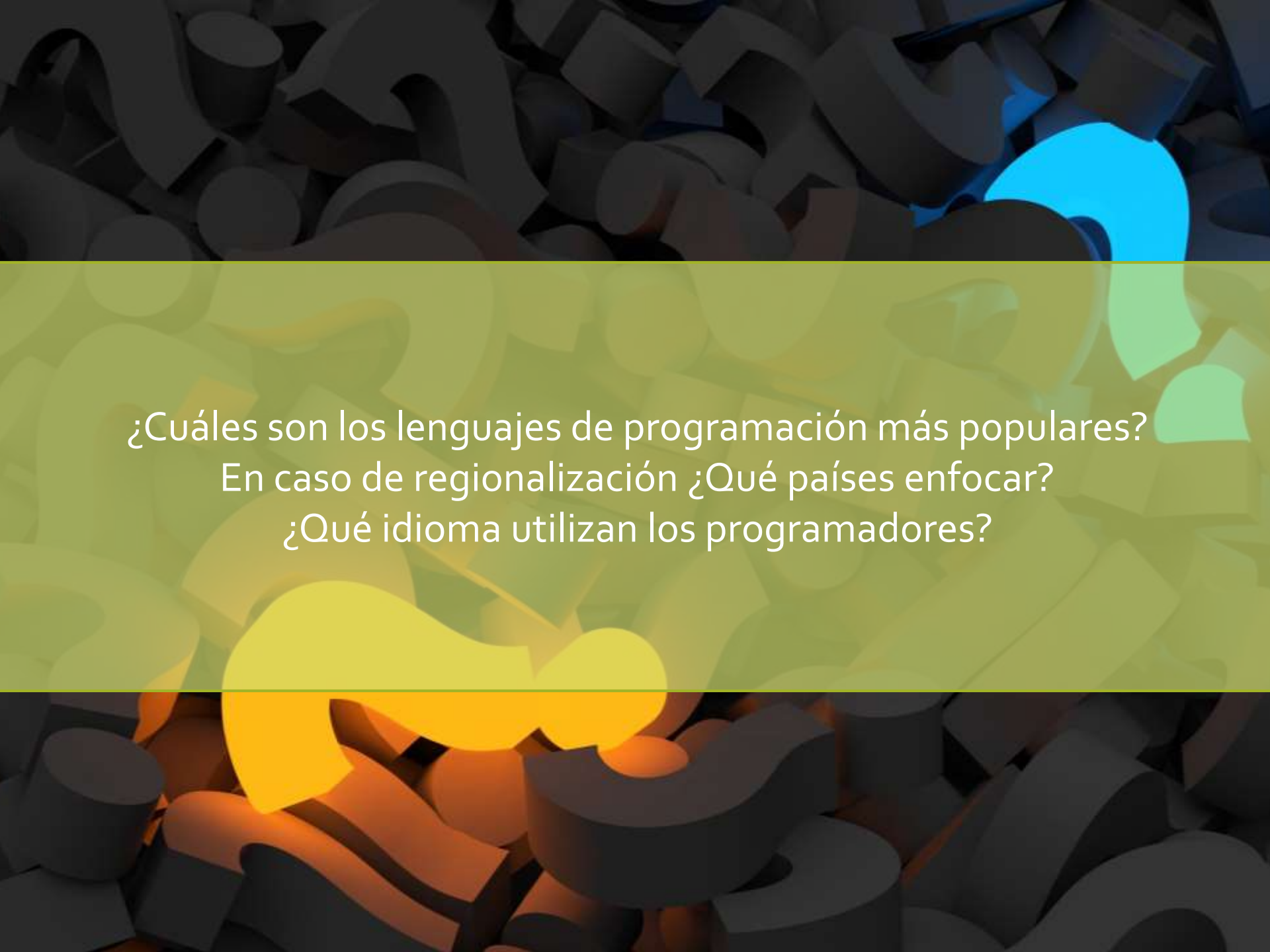
ID	Nombre	Fecha de nacimiento	Domicilio	Población	Código Postal
23	Pedro Gómez	10/12/1965	Avenida España, 36	Terrassa	8224

Big Data – Sistemas distribuidos

$$(\sqrt{144}) + (127/5) + (11^2) + (27 \times 42) + \left(\frac{1}{2} \times \frac{3}{2}\right)$$



PROYECTO



¿Cuáles son los lenguajes de programación más populares?
En caso de regionalización ¿Qué países enfocar?
¿Qué idioma utilizan los programadores?

Fuentes de datos

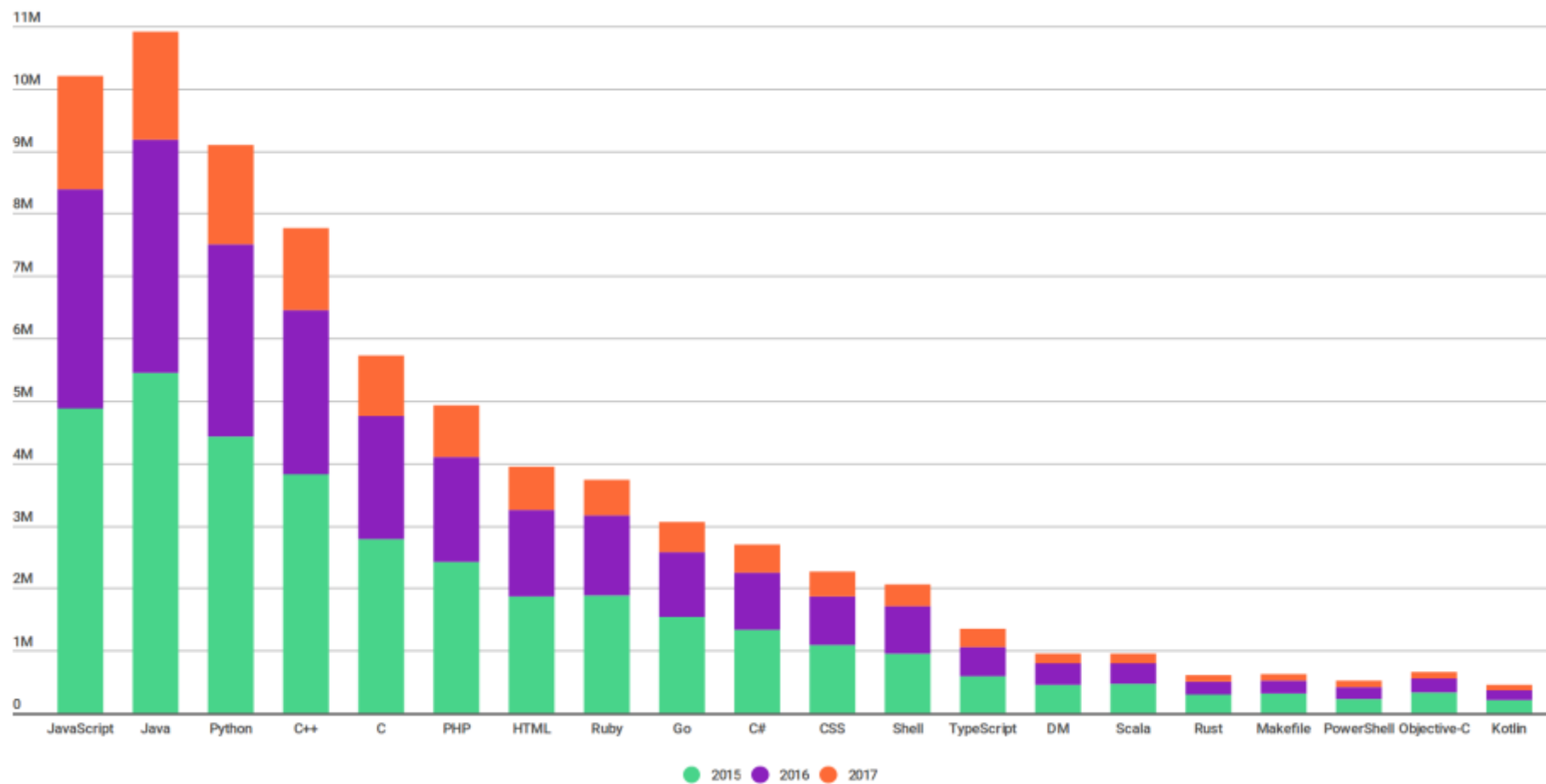


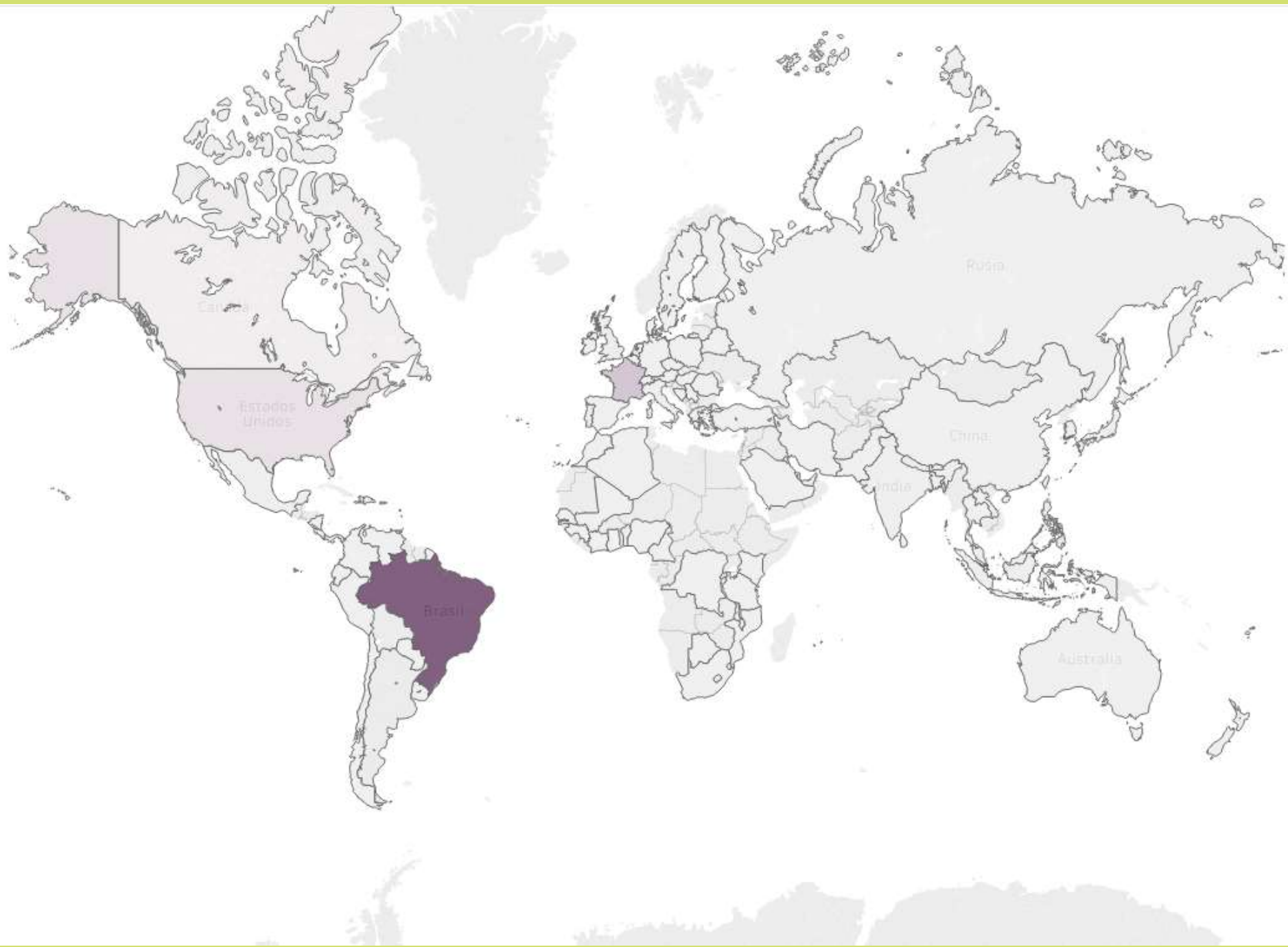
Cluster Hadoop



Insights







Idioma	Texto
Portugués	quanto mais converso C ge mais me conhecia antes C ve de novo agora, mais tenho ctz q eu era um lixo, jesus q ranço
Portugués	ir na leroy merlin e sair C a maior v... de... a própria casa
Portugués	Hj descobri q tem uma pessoa no ba... are amarela, ainda bem que tomei a vacina.
Francés	Ca me brise le coeur prcq C est... me de... force d'avancer et qui me fait sourire
Francés	C est quoi ce charabia de... C est quoi cette... ellerie ??
Francés	jsais pas comment... or d'enfants et d'les aimer m... suppose que C cque je ressent pr benjamin

Fail	Correct
Javascript, c#, c, c++, p, i, android, pyhon, java	javascript, androiddev , python , c++ , php , iosdev , html, kotlin, php, typescript, powershell, git, github, groovy, maven, cplusplus, springboot, haskell, androidstudio, intellij, netbeans, oracle, mysql, cprogramming, c#

