YO, CIENCIA DE DATOS

BIG DATA DAY
Facultad de Ciencias, UNAM
Ciudad de México, marzo 2016

Dr. Gabriel Guerrero
Socio Fundador y Director General
www.saxsa.com.mx

¿Qué es CIENCIA DE DATOS?

La **ciencia de datos** es un campo interdisciplinario que involucra los procesos y sistemas para extraer conocimiento o un mejor entendimiento de grandes volúmenes de datos.

Dr. Gabriel Guerrero
Socio Fundador y Director General
www.saxsa.com.mx

¿Qué es Big Data?

4V

Variedad:

Los datos pueden tener diferentes formas (estructurados o no estructurados) y formatos (.txt, .dat, .doc, .jpg, etcétera).

Volumen:

Los datos pueden manejar de enormes volúmenes

Velocidad:

Los resultados deben ser obtenidos con velocidad casi instantánea

Visibilidad:

Que los resultados sean expresados en forma visible a un usuario

Para resolver problemas **Big Data** (4V)
Se utiliza la **ciencia de datos**

Dr. Gabriel Guerrero
Socio Fundador y Director General
www.saxsa.com.mx

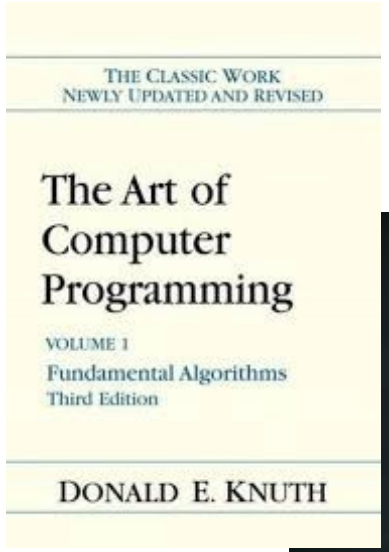
¿De dónde surge o en qué se basa?

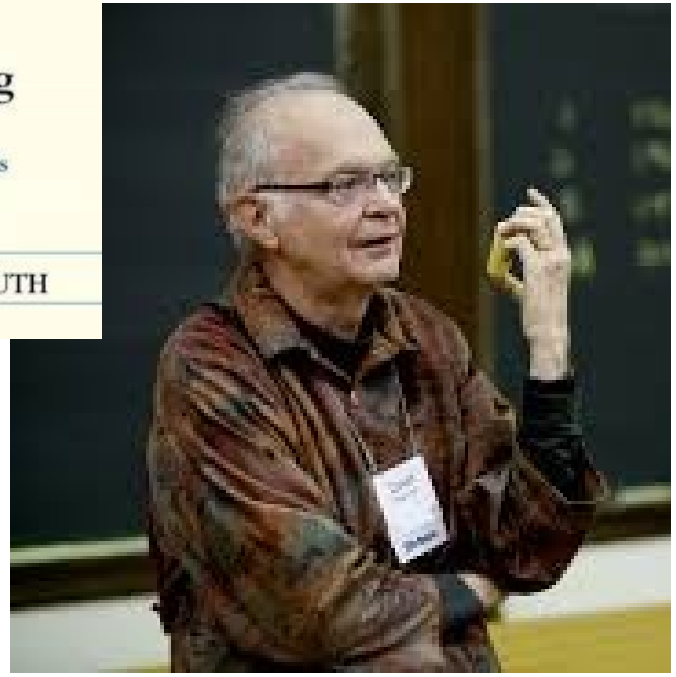
En técnicas e instrumentos de:

1. Ciencias de la computación
2. Teoría de la información
3. Estadística descriptiva e inferencial
5. Análisis de datos como:
   - Minería de datos (data mining)
   - Aprendizaje automatizado (machine learning)
6. Informática con lenguajes de programación (Java, Scala, Python, R)
7. Informática con métodos e instrumentos de cómputo distribuido y tolerante a fallas
8. HDFS (Hadoop Distributed File System)
9. Apache Spark
10. BDAS (Berkeley Data Analytics Stack)

Dr. Gabriel Guerrero
Socio Fundador y Director General
www.saxsa.com.mx

Grandes Precursores





Donald Ervin Knuth (1938),

Profesor emérito de la U de Stanford, es uno de los más reconocidos expertos en ciencias de la computación y pilar del análisis de algoritmos y técnicas de compilación

Dr. Gabriel Guerrero
Socio Fundador y Director General
www.saxsa.com.mx

Grandes Precursores

John E. Hopcroft (1939)

Algunos de sus libros:

- J.E. Hopcroft, Rajeev Motwani, Jeffrey D. Ullman, *Introduction to Automata Theory, Languages, and Computation* (2001)
- Alfred V. Aho, J.E. Hopcroft, Jeffrey D. Ullman, *Data Structures and Algorithms* (1983)
- Alfred V. Aho, J.E. Hopcroft, Jeffrey D. Ullman, *The Design and Analysis of Computer Algorithms* (1974).

Dr. Gabriel Guerrero
Socio Fundador y Director General
www.saxsa.com.mx

Herramientas actuales

Hoy se cuenta con una plataforma que ofrece un gran soporte a la **ciencia de datos.**

Esto es el conjunto de herramientas del BigData o técnicas para el manejo de grandes volúmenes de datos.

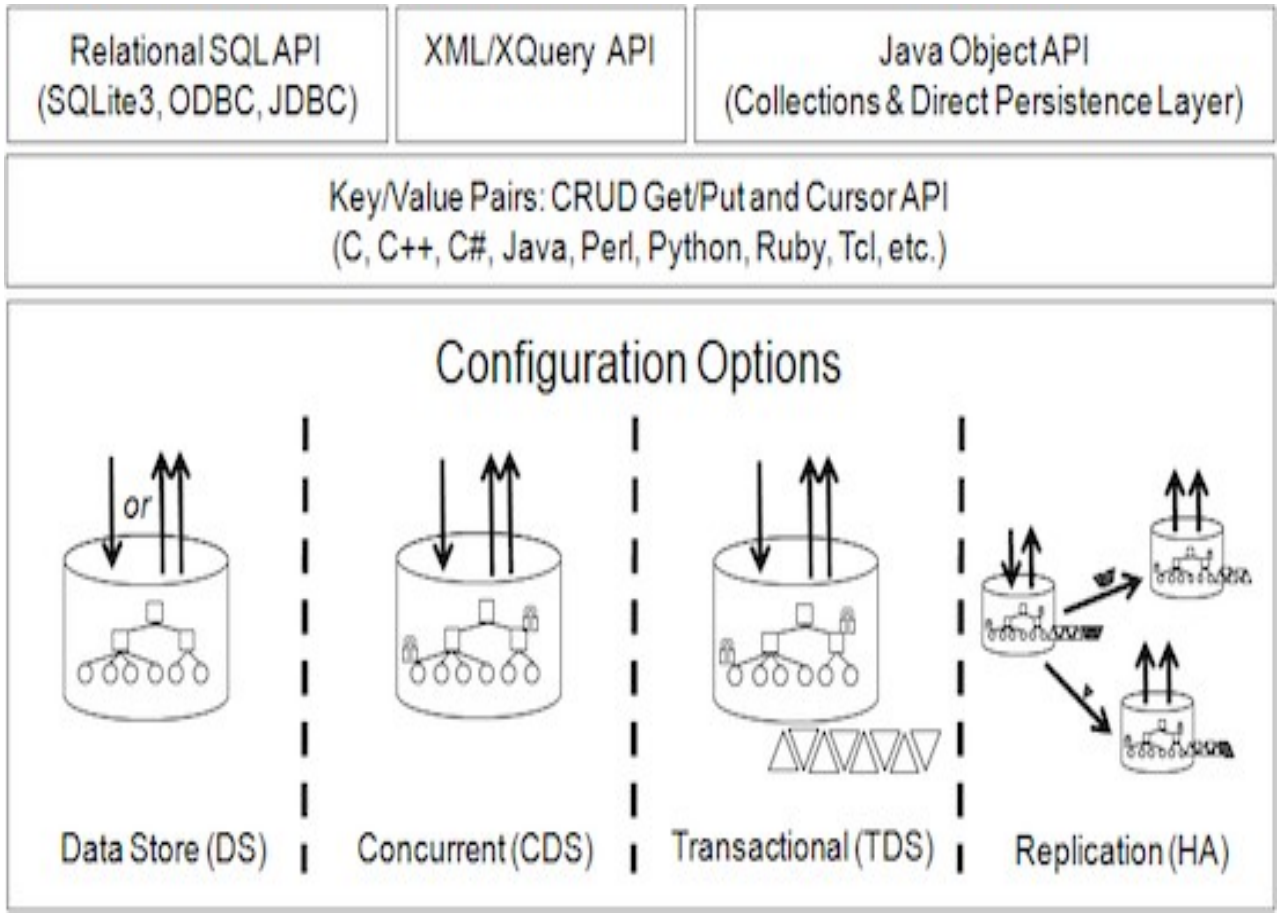
Dr. Gabriel Guerrero
Socio Fundador y Director General
www.saxsa.com.mx

Herramientas actuales

Tenemos entre otras:

1. Bibliotecas de árboles Btree, Berkeley DB (Sleepycat, comprada por Oracle en 2006)

http://www.oracle.com/technetwork/database/database-technologies/berkeleydb/overview/index.html



Dr. Gabriel Guerrero
Socio Fundador y Director General
www.saxsa.com.mx

Herramientas actuales

Tenemos entre otras:

2. Apache Hadoop HDFS/MapReduce

http://hadoop.apache.org/

Dr. Gabriel Guerrero
Socio Fundador y Director General
www.saxsa.com.mx

Herramientas actuales

BDAS, Berkeley Data Analytics Stack

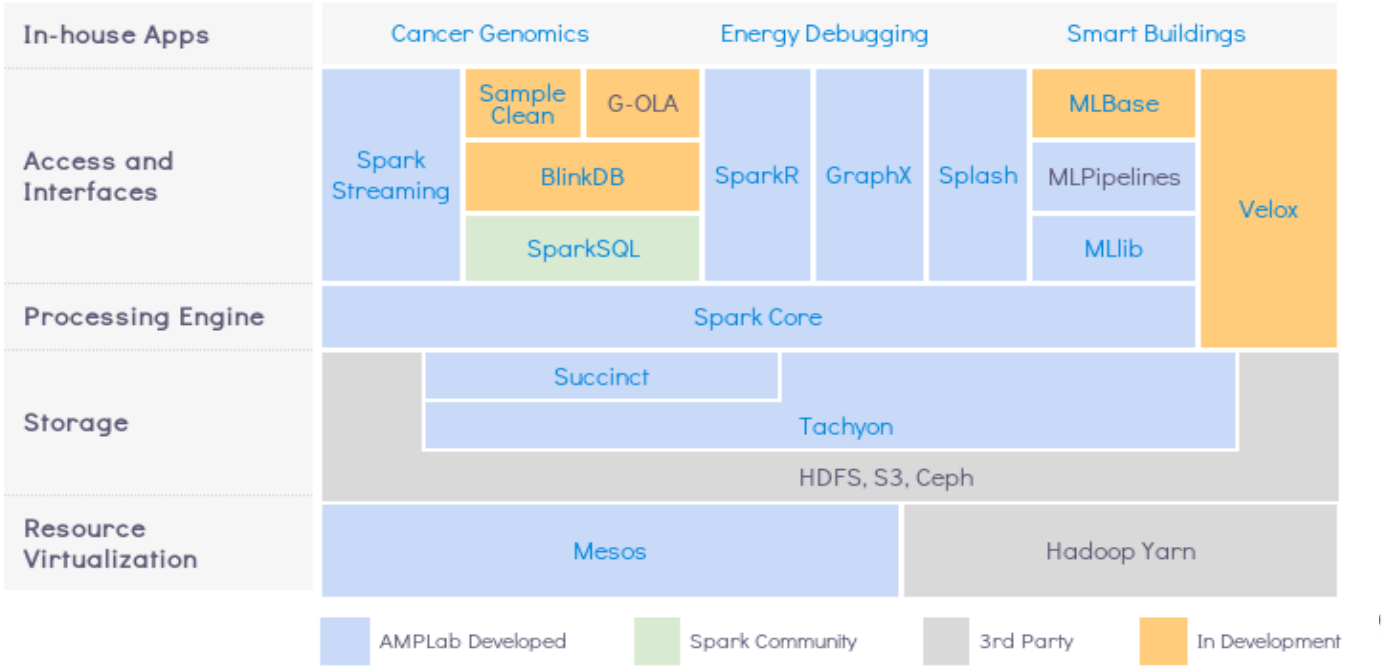
Con el objetivo de ofrecer un ambiente integral para el conjunto de aplicaciones de grandes volúmenes de datos distribuidos y tolerante a fallos, el marco de referencia (framework) Spark continúa añadiendo componentes en una forma unificada.

Este es el proyecto BDAS **(Berkeley Data Analytics Stack)**

Dr. Gabriel Guerrero
Socio Fundador y Director General
www.saxsa.com.mx

Herramientas actuales

BDAS es una pila de aplicaciones de código abierto (open source software stack) que integra las componentes de sistemas construidas por el Laboratorio AMPLab de UC Berkeley alrededor del concepto de Analíticos de Grandes Volúmenes (**BigData Analytics**)

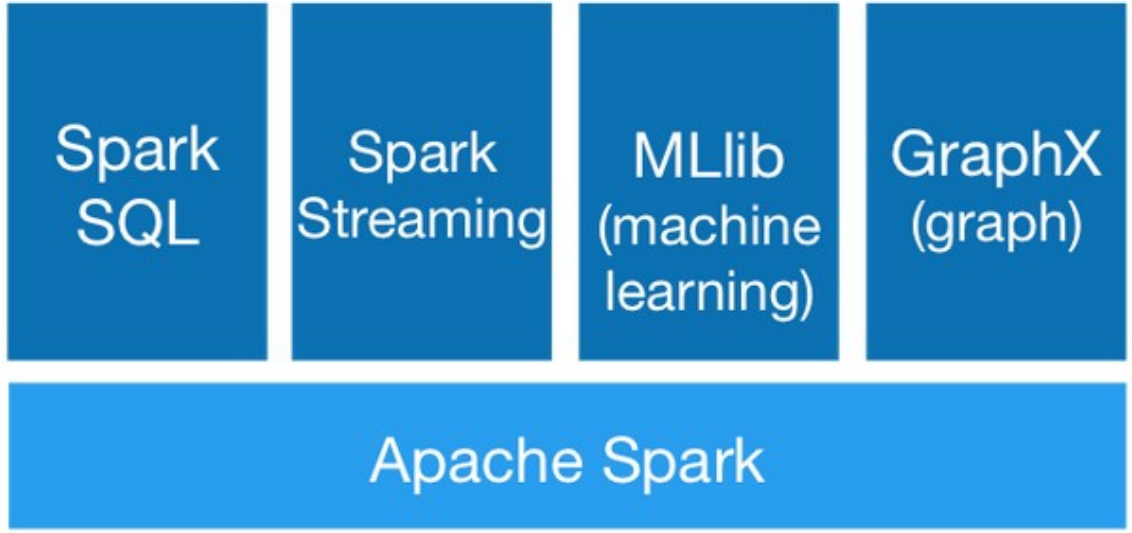


Dr. Gabriel Guerrero
Socio Fundador y Director General
www.saxsa.com.mx

Herramientas actuales

Apache Spark

El marco de referencia (framework) **Spark** ofrece un conjunto de bibliotecas para el desarrollo de aplicaciones. Entre las bibliotecas básicas integradas en la versión núcleo se tienen:



Dr. Gabriel Guerrero
Socio Fundador y Director General
www.saxsa.com.mx

Herramientas actuales

Spark SQL Es el modulo para el trabajo con datos estructurados.

Spark Streaming permite la construcción sencilla de aplicaciones escalables y tolerantes a fallas utilizando torrentes de datos.

MLlib es una biblioteca escalable con algoritmos de **Aprendizaje Automatizado** (scalable machine learning library).

GraphX es la biblioteca de Apache **Spark** para grafos y cómputo paralelo de grafos.

Dr. Gabriel Guerrero
Socio Fundador y Director General
www.saxsa.com.mx

Herramientas actuales

Algoritmos de Aprendizaje Automatizado (ML) en Spark

MLlib es la biblioteca de Spark que es escalable de algoritmos de Aprendizaje Automatizado (Machine Learning Library) que incluye herramientas como clasificación (classification), regresión (regression), agrupación (clustering), filtrado colaborativo (collaborative filtering), reducción de dimensiones (dimensionality reduction).

Dr. Gabriel Guerrero
Socio Fundador y Director General
www.saxsa.com.mx

Herramientas actuales

Hoy la idea es NO INVENTAR EL AGUA TIBIA en el Aprendizaje Automatizado de Grandes Volúmenes de Datos (BigData Analytics), y aplicar el tipo de algoritmo más conveniente a un problema y NO PROGRAMAR DESDE EL INICIO los algoritmos.

Hoy debemos pensar en los Algoritmos de Aprendizaje Automatizado como instrumentos o utensilios como en una analogía de un gran banquete

Dr. Gabriel Guerrero
Socio Fundador y Director General
www.saxsa.com.mx

Herramientas actuales

Hoy los comensales quieren bocadillos y banquetes JUSTO A TIEMPO y de FORMA INSTANTANEA (Streaming) para una gran población (usuarios Internet) en un extenso territorio (cobertura Internet), sin importar si en la cocina se prepararon con el mejor horno (Spark y BDAS) y las mejores recetas (Algoritmos ML eficientes) !!

Spark Streaming

Dr. Gabriel Guerrero
Socio Fundador y Director General
www.saxsa.com.mx

Herramientas actuales

Hoy **Spark** ofrece uno de los mejores instrumentos con una gran variedad de “recetas de ML”, en particular ya se cuenta en MLlib 1.6.1 los siguientes algoritmos listos para utilizarse:

Ver página Spark ML

http://spark.apache.org/docs/latest/mllib-guide.html

Dr. Gabriel Guerrero
Socio Fundador y Director General
www.saxsa.com.mx

Herramientas actuales

- Tipos de datos
- Estadísticas Básicas
  - Estadísticas de Resumen
  - Correlación
  - Muestreo estratificado
  - Prueba de hipótesis
  - Generación aleatoria de datos

- Clasificación y regresión
  - Modelos lineales (Máquinas de vectores de soporte [SVMs], regresión logística, regresión lineal)
  - Clasificador Bayesiano ingenuo
  - Árboles de decisión
  - Conjuntos de Árboles (Bosques Aleatorios y Árboles de decisión impulsados)
  - Regresión Isotónica

Dr. Gabriel Guerrero
Socio Fundador y Director General
www.saxsa.com.mx

Herramientas actuales

- Filtrado Colaborativo
  - Mínimos Cuadrados Alternantes (ALS)
- Agrupamiento
  - K-medias
  - Mezclas Gaussianas
  - Agrupamiento por método de las potencias (PIC)
  - Asignación Latente de Dirichlet (LDA)
  - Streaming de K-medias
- Reducción de dimensión
  - Descomposición de valor único (SVD)
  - Análisis de componentes principales (PCA)
- Extracción y transformación de características
- Minería por patrones frecuentes
  - Crecimiento FP
- Optimización (desarrollador)
  - Método de gradiente descendente estocástico
  - Límite de memoria BFGS (L-BFGS)
- Exportación de modelos PMML

Dr. Gabriel Guerrero
Socio Fundador y Director General
www.saxsa.com.mx

Compañías que recientemente han incorporado técnicas Spark

Dr. Gabriel Guerrero
Socio Fundador y Director General
www.saxsa.com.mx

Dr. Gabriel Guerrero
Socio Fundador y Director General
www.saxsa.com.mx