

Release 0.0.1

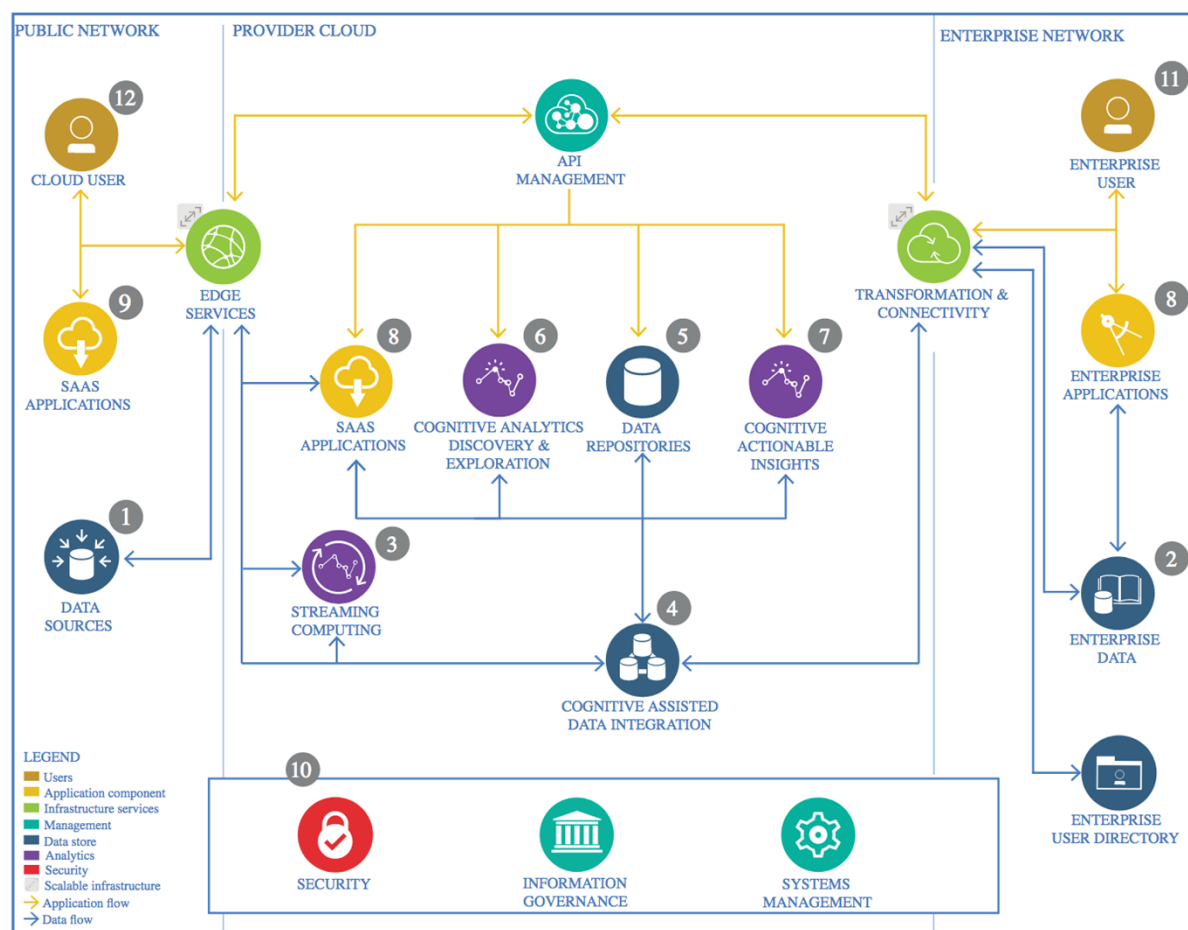
Date April 03, 2019

The datascience documentation was written by Asif Peshkar for IBM Advanced Datascience Capstone project from Coursera.

The Lightweight IBM Cloud Garage Method for Data Science

Architectural Decisions Document

1 Architectural Components Overview



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

1.1 Overview

A lot can be derived from buying trends. Future predictions can be made to create more availability of products to consumers. But, as a business provider, how can they predict exactly what needs to be stored, how much and when.

A great example can be through evaluation of past businesses. We make use of past records to provide insights for future sales. For this, we need a good source of data that can be explored and analysed.

1.2 Enterprise Data

Enterprise data is data that is shared by the users of an organization, generally across departments and/or geographic regions. Because enterprise data loss can result in significant financial losses for all parties involved, enterprises spend time and resources on careful and effective data modeling, solutions, security and storage. Enterprise data characteristics include:

- **Integration:** Ensures a single consistent version of enterprise data for sharing throughout an organization
- **Minimized redundancy, disparity and errors:** As enterprise data is shared by all of an organization's users, data redundancy and disparity must be minimized. Data modeling and management strategies are directed towards these requirements.
- **Quality:** To ensure data quality, enterprise data must follow organizational or other identified standards for varying internal and external data components.
- **Scalability:** Data must be scalable, flexible and robust to meet different enterprise requirements.
- **Security:** Enterprise data must be secured via authorized and controlled access.

1.3 Streaming analytics

Streaming Analytics is the ability to constantly calculate statistical **analytics** while moving within the **stream** of data. **Streaming Analytics** allows management, monitoring, and real-time **analytics** of live **streaming data**

Streaming Analytics involves knowing and acting upon events happening in your business at any given moment. Since Streaming Analytics occurs immediately, companies must act on the analytics data quickly within a small window of opportunity before the data loses its value.

1.4 Data Integration

Data integration is the combination of technical and business processes used to combine **data** from disparate sources into meaningful and valuable information. A complete **data integration** solution delivers trusted **data** from various sources

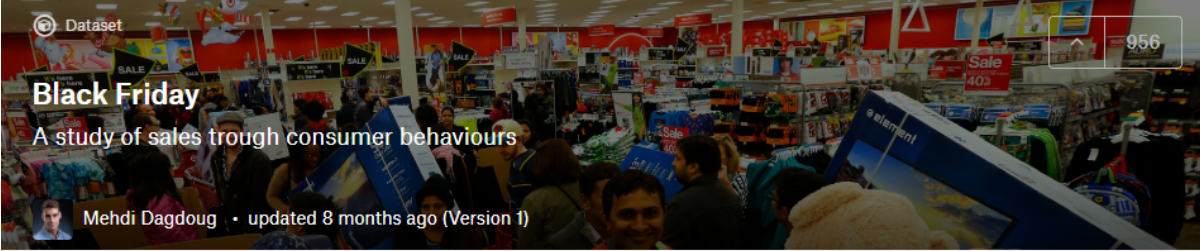
This process becomes significant in a variety of situations, which include both commercial (such as when two similar companies need to merge their databases) and scientific (combining research results from different bioinformatics repositories, for example) domains. Data integration appears with increasing frequency as the volume (that is, big data[2]) and the need to share existing data explodes.


1.5 Data Repository

Data repository is a somewhat general term used to refer to a destination designated for data storage.

For our example, we have used the Black Friday dataset from kaggle. This dataset is available at ...


<https://www.kaggle.com/mehdidag/black-friday>



 Dataset

Black Friday


A study of sales through consumer behaviours


 Mehdi Dagdoug • updated 8 months ago (Version 1)

[Data](#)
[Kernels \(158\)](#)
[Discussion \(19\)](#)
[Activity](#)
[Metadata](#)

Download (5 MB)

New Kernel

 CC0: Public Domain

 business, regression analysis

Description


Description

The dataset here is a sample of the transactions made in a retail store. The store wants to know better the customer purchase behaviour against different products. Specifically, here the problem is a regression problem where we are trying to predict the dependent variable (the amount of purchase) with the help of the information contained in the other variables.

Classification problem can also be settled in this dataset since several variables are categorical, and some other approaches could be "Predicting the age of the consumer" or even "Predict the category of goods bought". This dataset is also particularly convenient for clustering and maybe find different clusters of consumers within it.

Data (5 MB)









Data Sources

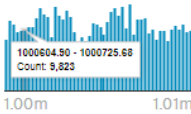
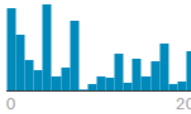
 BlackFriday.csv	538k x 12
---	-----------

About this file

Dataset of 550 000 observations about the black Friday in a retail store, it contains different kinds of variables either numerical or categorical. It contains missing values.

Columns

-  User_ID User
-  Product_ID Id Product
-  Gender Boolean
-  Age Age customer
-  Occupation Id Occupation of each customer
-  City_Category
-  Stay_In_Current_City_Years
-  Marital_Status

BlackFriday.csv (23.8 MB)							12 of 12 columns	Views						
	User_ID User	Product_ID Id Product	Gender Boolean	Age Age customer	Occupation Id Occupation of each customer	City								
		3623 unique values	M 75% F 25%	26-35 40% 36-45 20% Other (5) 40%		B C Other (1								
1	1000001	P00069042	F	0-17	10	A								
2	1000001	P00248942	F	0-17	10	A								
3	1000001	P00087842	F	0-17	10	A								
4	1000001	P00085442	F	0-17	10	A								
5	1000002	P00285442	M	55+	16	C								
6	1000003	P00193542	M	26-35	15	A								
7	1000004	P00184942	M	46-50	7	B								
8	1000004	P00346142	M	46-50	7	B								
9	1000004	P0097242	M	46-50	7	B								
10	1000005	P00274942	M	26-35	20	A								
11	1000005	P00251242	M	26-35	20	A								
12	1000005	P00014542	M	26-35	20	A								
13	1000005	P00031342	M	26-35	20	A								
14	1000005	P00145042	M	26-35	20	A								

1.5.1 Technology Choice

The dataset collected from Kaggle has been uploaded at Asif Peshkar's IBM Watson Studio repository as an asset. The dataset has been named as BlackFriday.csv, all future revisions to this dataset should replace this original file keeping the name constant.

IBM Watson Studio

Upgrade

Asif Peshkar's Account

AP

My Projects / AdvancedDataScience

Launch IDE

Add to project

Overview

Assets

Environments

Bookmarks

Deployments

Access Control


Settings

What assets are you looking for?

New data asset

▼ Data assets

0 asset selected.

<input type="checkbox"/>	NAME	TYPE	CREATED BY	LAST MODIFIED	ACTIONS
<input type="checkbox"/>	 BlackFriday.csv	Data Asset	Asif Peshkar	1 Apr 2019, 7:46:02 pm	

1.6 Discovery and Exploration

A Python 3 notebook has been used for discovery and exploration of data. All insights and data presentation is done in this notebook.

The notebook can be accessed through this notebook with write/update privileges to collaborators...

https://dataplatform.cloud.ibm.com/analytics/notebooks/v2/a06de59f-ef04-4361-9365-2b5714001cc4/view?access_token=1c0d1420a73f5278fe9954f84482f3dd3ed14e4472f61a56739aa2cb265c3cdb

1.6.1 Technology Choice

Data handling is done using numpy and pandas framework.

Data analysis is done using Seaborn, matplotlib, and pyplot.

Data modelling is done using **scikit-learn library**

1.7 Actionable Insights

Various insights have been produced and explained in the document after data analysis and modeling.

1.8 Applications / Data Products

Results and Conclusions have been derived using the findings. Results extracted from the notebook can be revealed for presentations.