Big Data Processing

Power by Open Source Data Cloud Lab

Portfolio of Data Analyst skill set

& Fellow member

BANGLADESH COMPUTER SOCIETY

Association for Computing Machinery

CIPS — Canada's Association of I.T. Professionals

1985

Research Assistant,
-Dhaka University and National Academy of Science, D.C., USA
-National Institute of Population Research and Training (NIPORT)
-The Department of Urban and Regional Planning at BUET
SW: dBase III Plus, Lotus, SPSS/PC+, Harvard Graphics

Data Processor and Analyst, Data Computerization Project,
-National Board of Revenue- The World Bank
SW: Paradox RDMS, dBase III Plus, Lotus, SPSS/PC+, Harvard

1995

Project Director, a research project as granted by
- The Ministry of Science Information & Communications
Technology, Govt. of Bangladesh, Medical SW for public health.
SW: MS Access, Excel , Micro Media Flash and Web Application

Datacenter CCBS Manager, CDR (Called Details Record) , Telco **Robi**
in 1997 as TMI (**Bangladesh**) with the brand name **AKTEL GSM'**.
SW: Sybase , Oracle RDMS, Excel

2000

Head of Telecom CCBS Application, CDR (Called Details Record) ,
Telebarta Limited (Jubok Phone) –Motorola Project
SW: Microsoft SQL Server, Oracle, Crystal Report

2010

Data Analyst and IT support
-SYNAP IT Corporation, New York, NY
-Document Technologies, Inc., Philadelphia, PA
-ISP, Food Bank,YSCP Charter School, State College, PA
-Micro Datacenter and Data Warehouse, Round Rock, TX.
SW: Microsoft SQL Server, Oracle,

Big Data and Data Warehouse Analyst and Hadoop Open Source
-Micro Datacenter and Data Warehouse
SW: Linux, Hadoop, HiveQL, Spark, Sqoop, Habase

2020

Data Cloud Lab
1 Million Data

Power by
AWS Cloud EMR, EC2 and S3, Cloudera & Core HiveQL

1M Data

Data Set:
1. Popular_Baby_Names Data_20000
2. Retails Data_250000
3. Olympic_Data_8618
4. Weblog_250
5. HealthCare_50000
6. Weather_Data_10000
7. TruckDrive_18878

Web, App Cloud, Database, Excel, Notepad, etc.

Data Sourc

1 Data source is the primary location from where data comes. The data source can be a database, a dataset, a spreadsheet or even hard-coded data

Excel/CVS/Text File Format:
John,30,New York
Robert,25,DC

Data Lake

2 A data lake is usually a single store of all enterprise data

JSON, Web Application File Format : {
"employee":{
"name":"John", "age":30, "city":"New York" }
}

Text File/Sequence File
RC File/ORC File
AVRO File
Parquet/Binary File

sqoop import-all-tables

Data Pipeline

3 a pipeline, also known as a data pipeline, is a set of data processing elements connected in series

Data Store

4 A *Data Store* is a connection to a store of data, whether the data is stored in a database or in one or more files

Linux O/S ,
Hadoop Big Data,
Hadoop File System(HDFS)
3 replication Datanode

Database
Customer
Product

Hive Database,
Data Warehouse

Data Warehouse

5 A data warehouse is a system that pulls together data from many different sources

Data Processing

6 Data processing is, generally, "the collection and manipulation of items of data to produce meaningful information

SELECT * FROM Customers;

SELECT Country FROM Customers;

SELECT * FROM Customers
WHERE Country='Mexico';

-Top 10 Revenue-generating products
-Total revenue per product
- Top 10 Diseases
-Weather analysis

Data Revenue

7 The statistic shows the *revenue* from the global big *data*

Data Visulization

8 Data visualization is the graphical representation of information and data

Version: v.001.2020