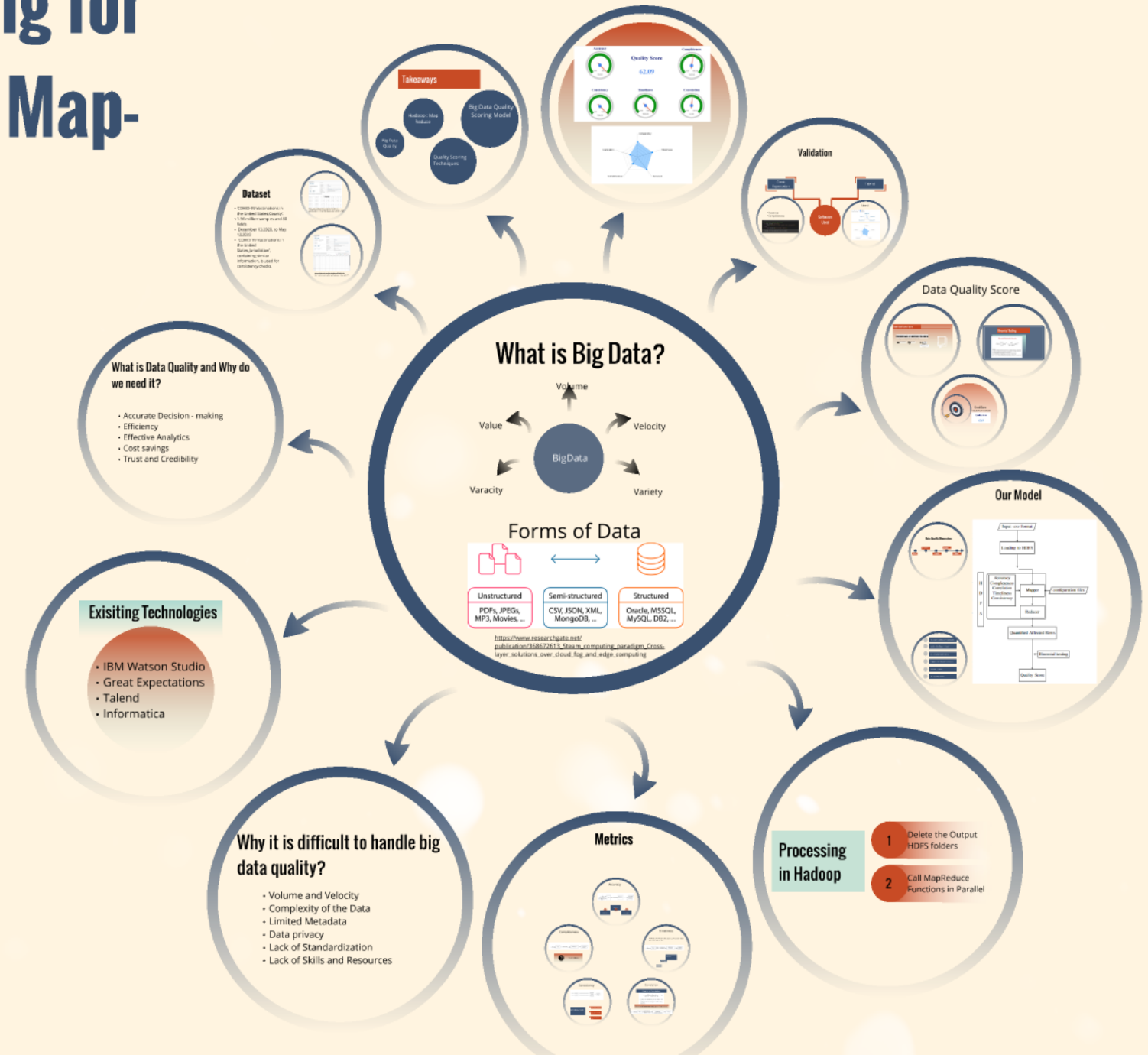
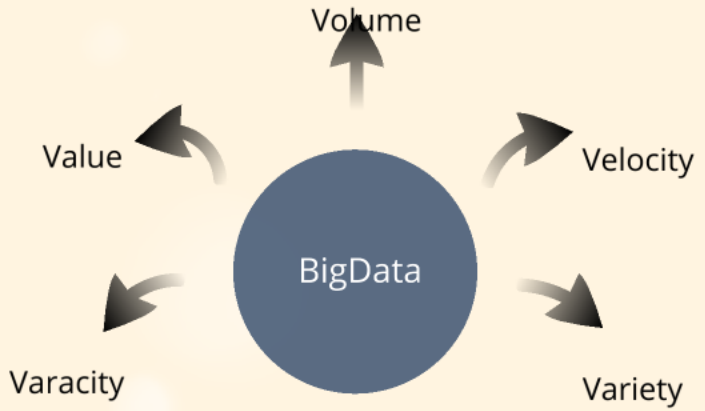
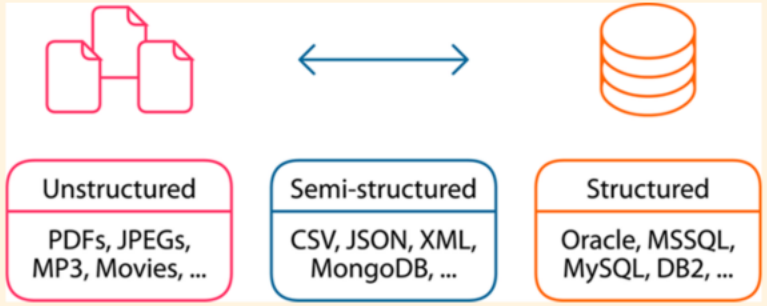
Big Data Quality Scoring for Structured Data Using Map-Reduce



What is Big Data?



Forms of Data



https://www.researchgate.net/publication/368672613_Steam_computing_paradigm_Cross-layer_solutions_over_cloud_fog_and_edge_computing

What is Data Quality and Why do we need it?

- Accurate Decision - making
- Efficiency
- Effective Analytics
- Cost savings
- Trust and Credibility

Why it is difficult to handle big data quality?

- Volume and Velocity
- Complexity of the Data
- Limited Metadata
- Data privacy
- Lack of Standardization
- Lack of Skills and Resources

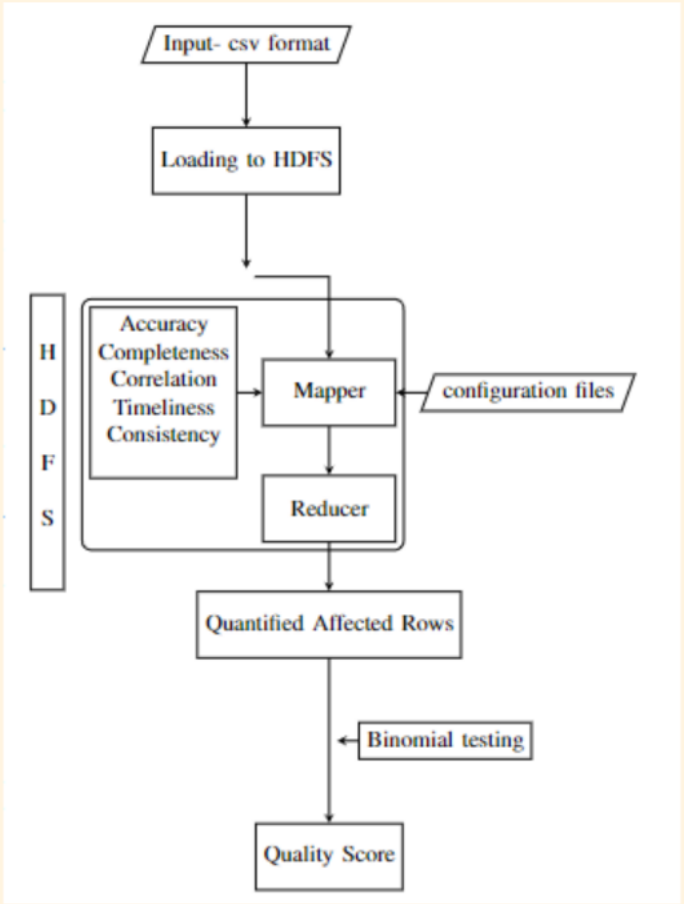
Exisiting Technologies

- IBM Watson Studio
- Great Expectations
- Talend
- Informatica

Our Model


Data Quality Dimensions


Hadoop MapReduce Framework
Input - User Requirement
Data Quality Dimensions
Mapper and reducer Functions
Binomial Testing
Data Quality Score


Input- csv format
Loading to HDFS
H
D
F
S
Accuracy
Completeness
Correlation
Timeliness
Consistency
Mapper
configuration files
Reducer
Quantified Affected Rows
Binomial testing
Quality Score

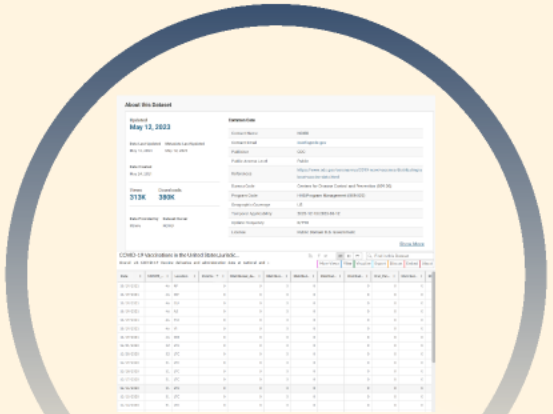

Hadoop MapReduce Framework
Input - User Requirement
Data Quality Dimensions
Mapper and reducer Functions
Binomial Testing
Data Quality Score


Data Quality Dimensions
Accuracy
Completeness
Consistency
Timeliness
Correlation

Dataset

- ‘COVID-19 Vaccinations in the United States,County’.
- 1.96 million samples and 80 fields
- December 13,2020, to May 12,2023
- ‘COVID-19 Vaccinations in the United States,Jurisdiction’, containing similar information, is used for consistency checks.

https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-County/8xkx-amqh



https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-Jurisdi/unsk-b7fc

n

y'.

d 80

About this Dataset

Updated
May 12, 2023

Data Last Updated
May 12, 2023

Metadata Last Updated
May 12, 2023

Date Created
May 24, 2021

Views
458K

Downloads
211K

Common Core

Contact Name	NCIRD
Contact Email	iisinfo@cdc.gov
Publisher	CDC
Public Access Level	Public
References	https://www.cdc.gov/coronavirus/2019-ncov/vaccines/distributing/about-vaccine-data.html
Bureau Code	Centers for Disease Control and Prevention (009:20)
Program Code	HHS:Program Management (009:020)
Geographic Coverage	US

COVID-19 Vaccinations in the United States,County

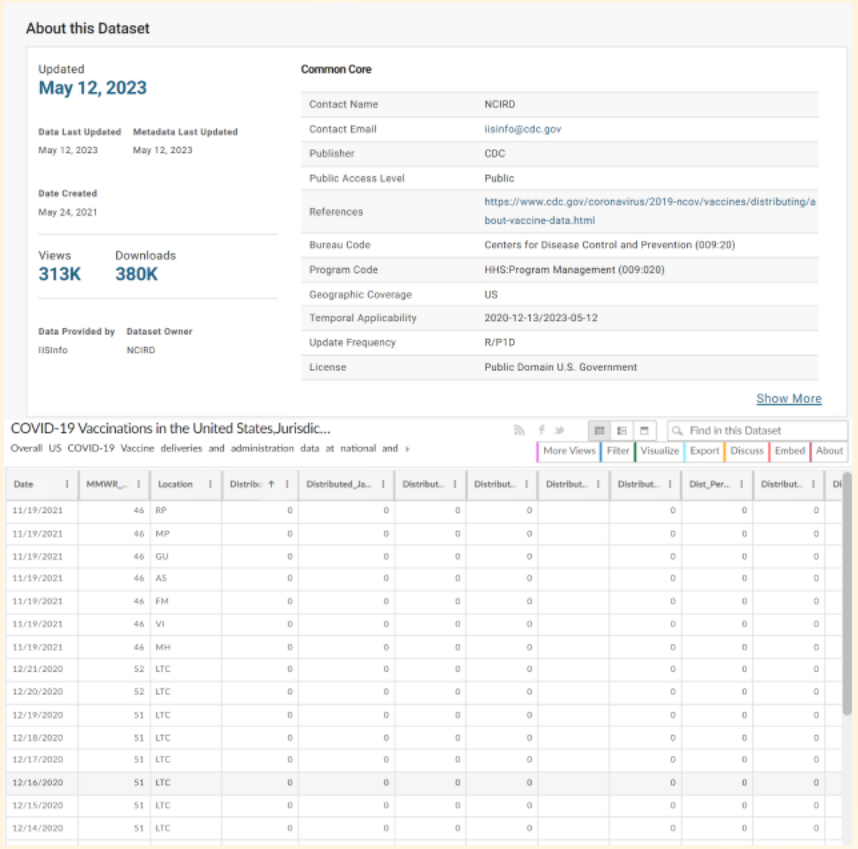
Date	FIPS	MMW...	Recip...	Recip...	Comp...	Admi...	Admi...	Admi...	Admi...	Admi...	Admi...	Admi...	Admi...	Admi...
							Administered_Dose1_Recip							
05/10/2023	55129	19	Washburn (	WI	96.7	11,123	70.8	11,097	73.9	10,863	78.5	10,368	81.3	
05/10/2023	19173	19	Taylor Cour	IA	97.3	3,149	51.4	3,145	55	3,079	59.8	2,966	63.3	
05/10/2023	36059	19	Nassau Cou	NY	97.5	1,391,226	95	1,384,503	95	1,329,779	95	1,232,671	95	
05/10/2023	48281	19	Lampasas (	TX	98.9	11,678	54.5	11,660	57.4	11,468	61.9	10,849	64.6	
05/10/2023	26145	19	Saginaw Co	MI	94	104,075	54.6							
05/10/2023	13101	19	Echols Cour	GA	88.8	1,573	39.3	1,572	41.4	1,537	46.6	1,458	49.2	
05/10/2023	12013	19	Calhoun Co	FL	98.6	6,023	42.7	6,016	44.8	5,971	48.5	5,753	50.9	
05/10/2023	17027	19	Clinton Cou	IL	97.7	24,993	66.5	24,936	70.5	24,266	75	22,777	76.8	

https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-County/8xkx-amqh

o May

ns in

or



About this Dataset

Updated
May 12, 2023

Data Last Updated	Metadata Last Updated
May 12, 2023	May 12, 2023

Date Created
May 24, 2021

Views	Downloads
313K	380K

Data Provided by	Dataset Owner
IISInfo	NCIRD

Common Core

Contact Name	NCIRD
Contact Email	iisinfo@cdc.gov
Publisher	CDC
Public Access Level	Public
References	https://www.cdc.gov/coronavirus/2019-ncov/vaccines/distributing/about-vaccine-data.html
Bureau Code	Centers for Disease Control and Prevention (009:20)
Program Code	HHS:Program Management (009:020)
Geographic Coverage	US
Temporal Applicability	2020-12-13/2023-05-12
Update Frequency	R/P1D
License	Public Domain U.S. Government

Show More

COVID-19 Vaccinations in the United States,Jurisdic...

Overall US COVID-19 Vaccine deliveries and administration data at national and ›

Find in this Dataset

More Views | Filter | Visualize | Export | Discuss | Embed | About

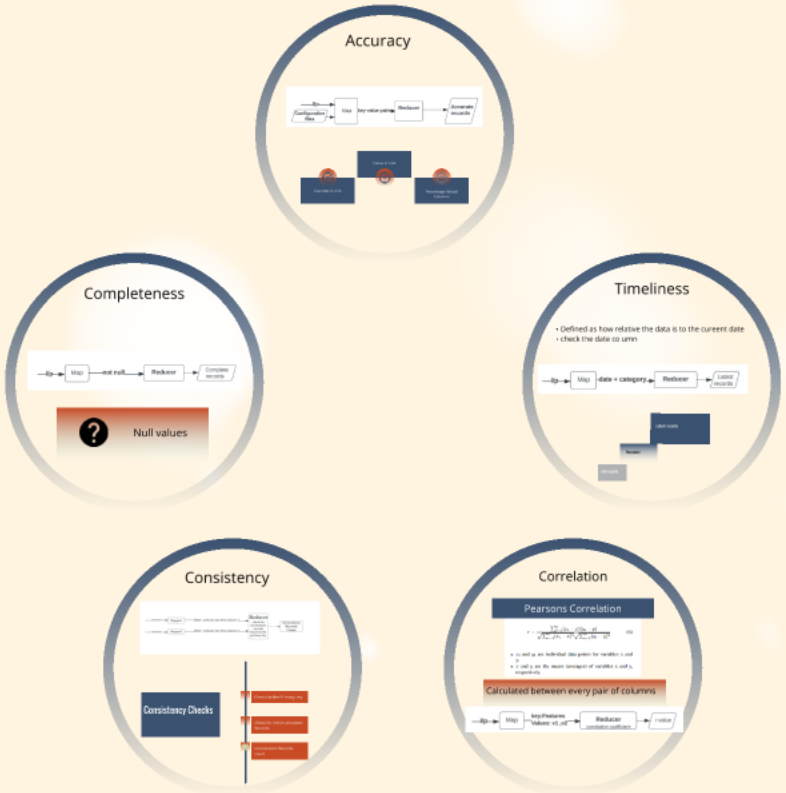
Date	MMWR_...	Location	Distrib...	Distributed_Ja...	Distribut...	Distribut...	Distribut...	Distribut...	Dist_Per...	Distribut...	Di
11/19/2021	46	RP	0	0	0	0		0	0	0	
11/19/2021	46	MP	0	0	0	0		0	0	0	
11/19/2021	46	GU	0	0	0	0		0	0	0	
11/19/2021	46	AS	0	0	0	0		0	0	0	
11/19/2021	46	FM	0	0	0	0		0	0	0	
11/19/2021	46	VI	0	0	0	0		0	0	0	
11/19/2021	46	MH	0	0	0	0		0	0	0	
12/21/2020	52	LTC	0	0	0	0		0	0	0	
12/20/2020	52	LTC	0	0	0	0		0	0	0	
12/19/2020	51	LTC	0	0	0	0		0	0	0	
12/18/2020	51	LTC	0	0	0	0		0	0	0	
12/17/2020	51	LTC	0	0	0	0		0	0	0	
12/16/2020	51	LTC	0	0	0	0		0	0	0	
12/15/2020	51	LTC	0	0	0	0		0	0	0	
12/14/2020	51	LTC	0	0	0	0		0	0	0	

https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-Jurisdi/unsk-b7fc

handle big

rces

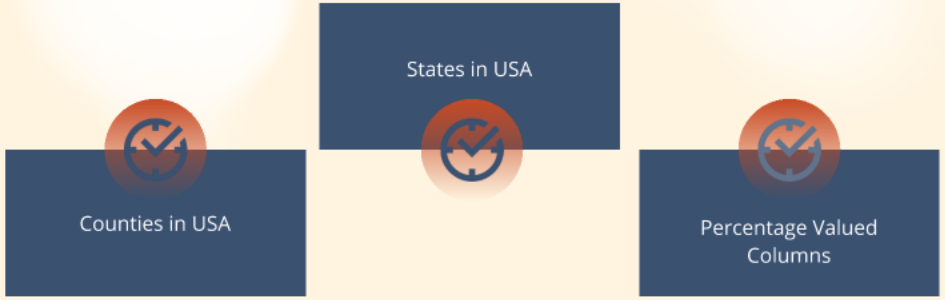
Metrics



Processing in Hadoop

Accuracy

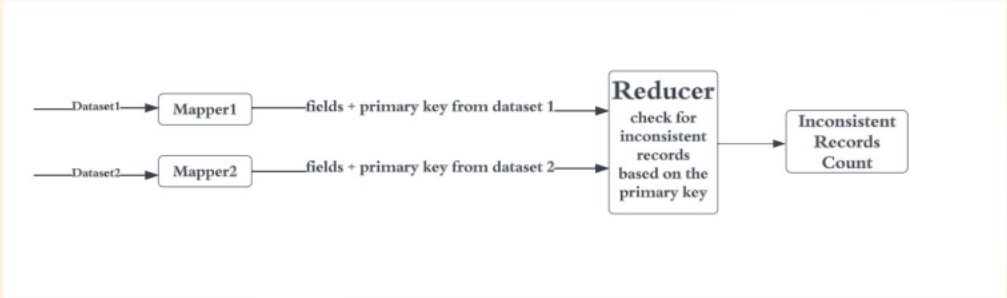

I/p
Configuration files
Map
key-value pairs
Reducer
Accurate records


States in USA
Counties in USA
Percentage Valued Columns

Completeness

Null values

Consistency

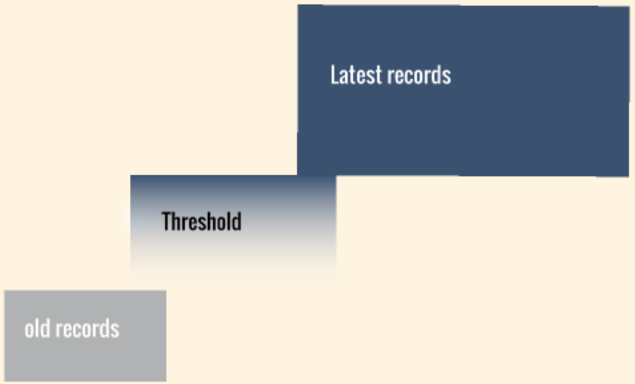


Consistency Checks

- Check for the Primary key
- Check for the Inconsistent Records
- Inconsistent Records count

Timeliness

- Defined as how relative the data is to the cureent date
- check the date column



Correlation

Pearsons Correlation

$$r = \frac{\sum_{i=1}^{n}(x_i - x)(y_i - y)}{\sqrt{\sum_{i=1}^{n}(x_i - x)^2}\sqrt{\sum_{i=1}^{n}(y_i - y)^2}} \quad (6)$$

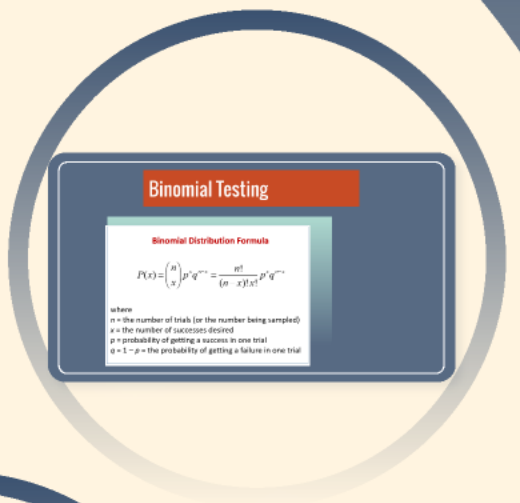
- x_i and y_i are individual data points for variables x and y.
- x and y are the means (averages) of variables x and y, respectively.

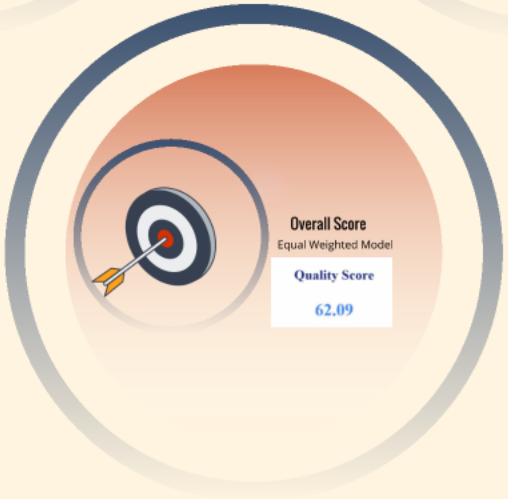
Calculated between every pair of columns

Processing in Hadoop

1. Delete the Output HDFS folders
2. Call MapReduce Functions in Parallel

Data Quality Score





Individual Feature Score

PERCENTAGE OF INEFFECTIVE ROWS

Calculating the percentage of ineffective rows in the dataset.

INEFFECTIVE ROWS

TOTAL ROWS

PERCENTAGE CALCULATION

$$\frac{\#ineffective\_rows}{total\_rows}$$

PERCENTAGE OF INEFFECTIVE ROWS

Calculating the percentage of ineffective rows in the dataset.

INEFFECTIVE ROWS

TOTAL ROWS

PERCENTAGE CALCULATION

$$\frac{\#ineffective\_rows}{total\_rows} \quad (1)$$

Binomial Testing

Binomial Distribution Formula

$$P(x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{(n-x)!\,x!} p^x q^{n-x}$$

where
n = the number of trials (or the number being sampled)
x = the number of successes desired
p = probability of getting a success in one trial
$q = 1 - p$ = the probability of getting a failure in one trial

Overall Score

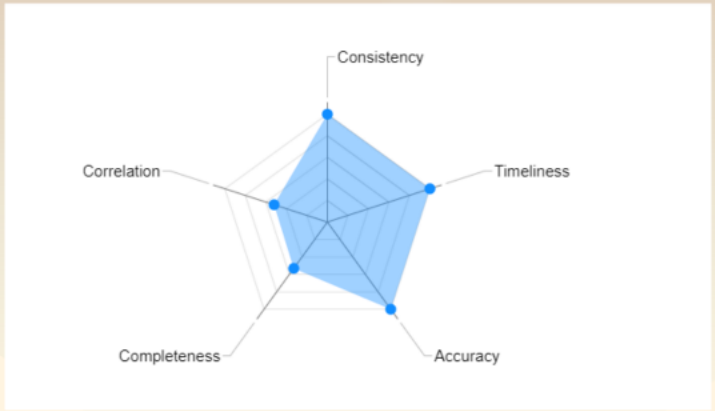
Equal Weighted Model

Quality Score

62.09

Big Data Quality Scoring Model

ng



Validatio

Great Expectation

Validation

Great Expectation

Talend

Softwares Used

- Accuracy
- Completeness

```
[ ] # Take action based on the results
    if test_results["success"]:
        print ("Awesome. All Data Quality Tests are green!")
    else:
        raise Exception("You've got issues.")
```

```
Awesome. All Data Quality Tests are green!
```

Talend

- Accuracy
- Completeness

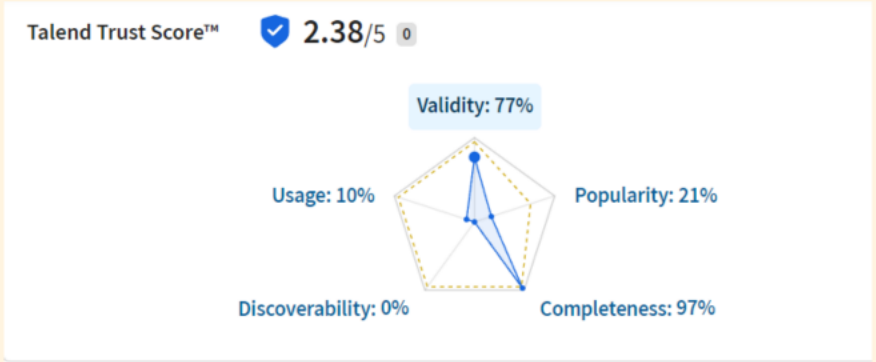
```
[ ]  # Take action based on the results
     if test_results["success"]:
         print ("Awesome. All Data Quality Tests are green!")
     else:
         raise Exception("You've got issues.")
```

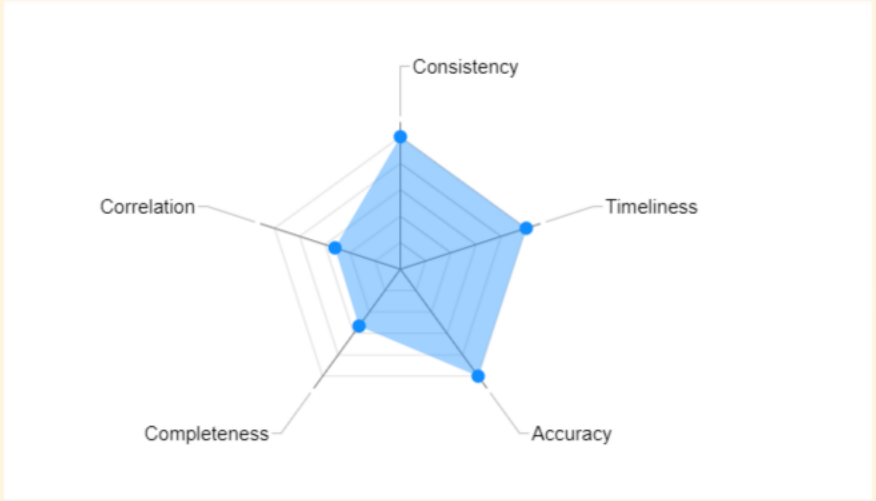
```
Awesome. All Data Quality Tests are green!
```

Soft
U

ftwares
Used

Talend





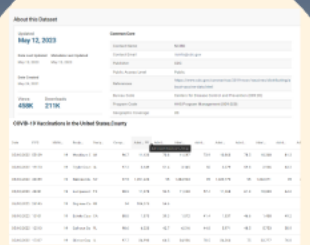
Takeaways

Hadoop : Map Reduce

Big Data Quality Scoring Model

Big Data Quality

Quality Scoring Techniques



https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-County/8xkx-amqh

Big Data Quality Scoring for Structured Data Using Map-Reduce