Data science: Case study

Problem statement

- **Big data**：The volume of this dataset is around 43GB. This is make it big data problem.
  - Chest X-Ray dataset has 112120 gray scale images with 1024 by 1024 image size.
- Multiple class with multiple labels:
  - The dataset contains x-ray images that shows one or more Thorax Disease and the total number of diseases is 14.
  - This make the problem as multiple class and multiple labels problem.

Challenges

- Big data : with 42 GB of images
- Data types : images and text thus NLP is required to extract labels
- Multiple class and multiple labels

❑Requirements
powerful hardware : Hadoop

- Project management tools
  - Confluence
  - JIRA
  - BITBUCKET

How Data are store

- Data Description:
  - Stored in CSV file which contains information about each patients and what diseases he/she has. This shown here
  - xray_data_statistics+.html
- Most important Columns :
  - Image index : unique name of each image
  - Finding Labels: List of all possible disease in each X-ray image
- Where is the labels :
  - NLP is used to extract the labels from the finding labels.
  - Convert_images_final_jan29.html
- Images are stored in sperate folder with unique names

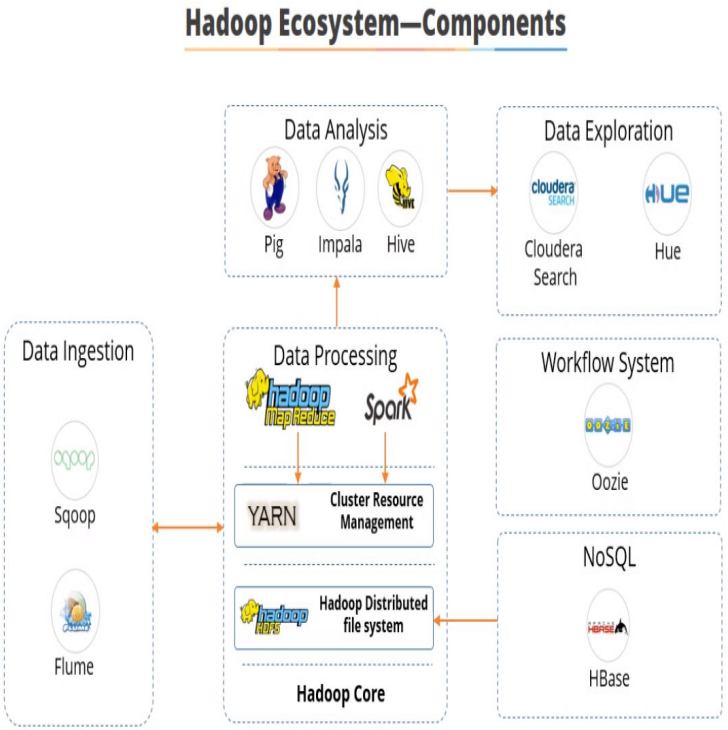
Dataset : Classes distribution

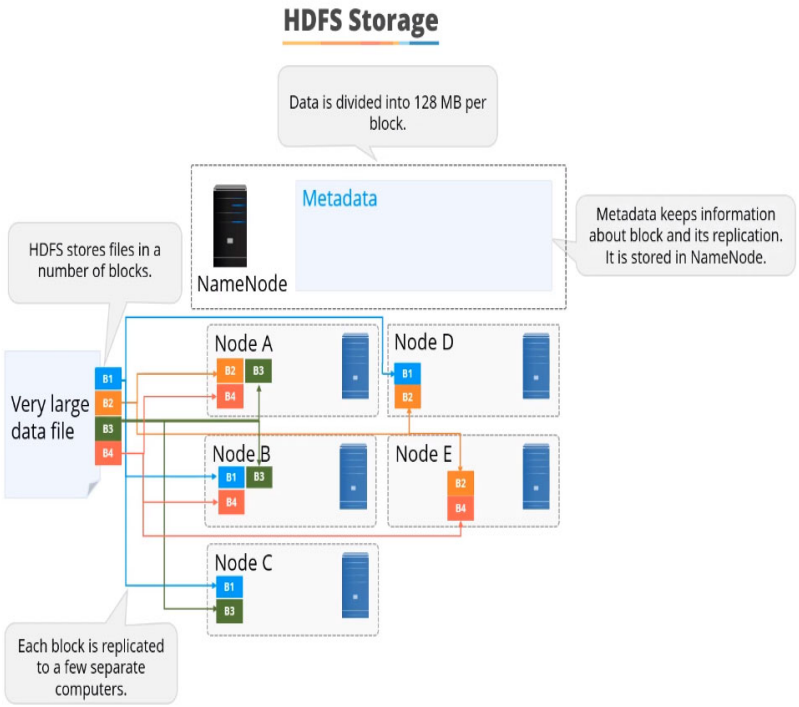
TABLE I
CHEST X-RAY CLASS DISTRIBUTION

Class Name	*Frequency*	*Percentage*
Atelectasis	11559.0	10.309490
Cardiomegaly	2776.0	2.475919
Effusion	13317.0	11.877453
Infiltration	19894.0	17.743489
Mass	5782.0	5.156975
Nodule	6331.0	5.646629
Pneumonia	1431.0	1.276311
Pneumothorax	5302.0	4.728862
Consolidation	4667.0	4.162504
Edema	2303.0	2.054049
Emphysema	2516.0	2.244024
Fibrosis	1686.0	1.503746
Pleural Thickening	3385.0	3.019087
Hernia	227.0	0.202462
No Finding	60361.0	53.836068

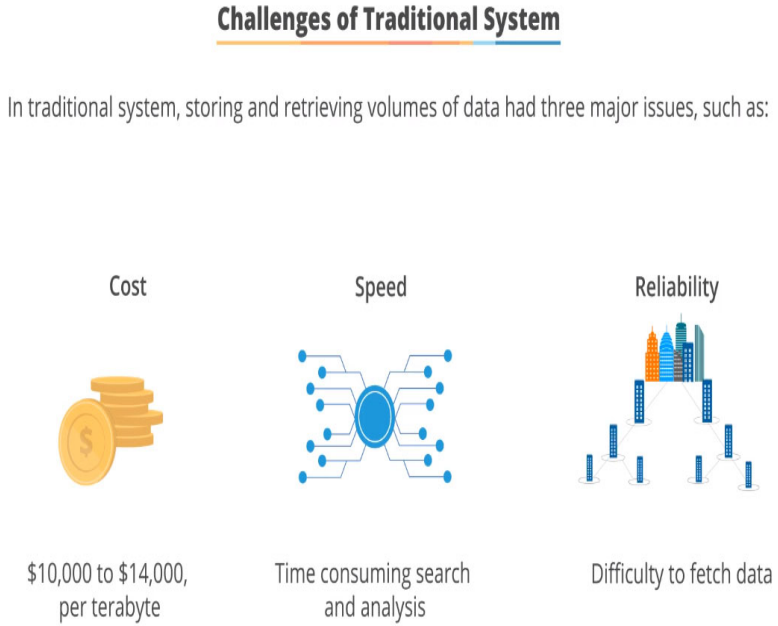
Imbalanced dataset

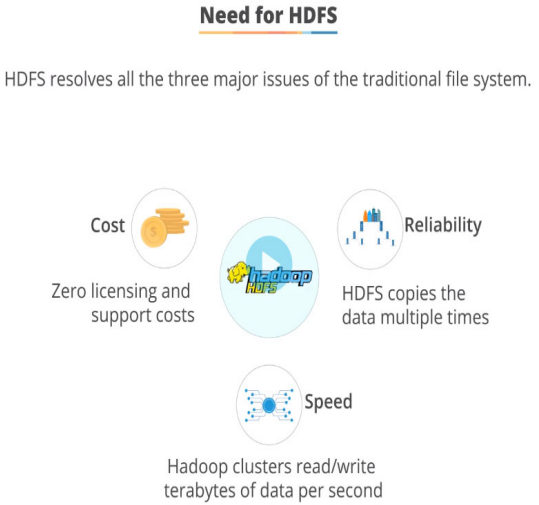
Requirements: Hardware infrastructure



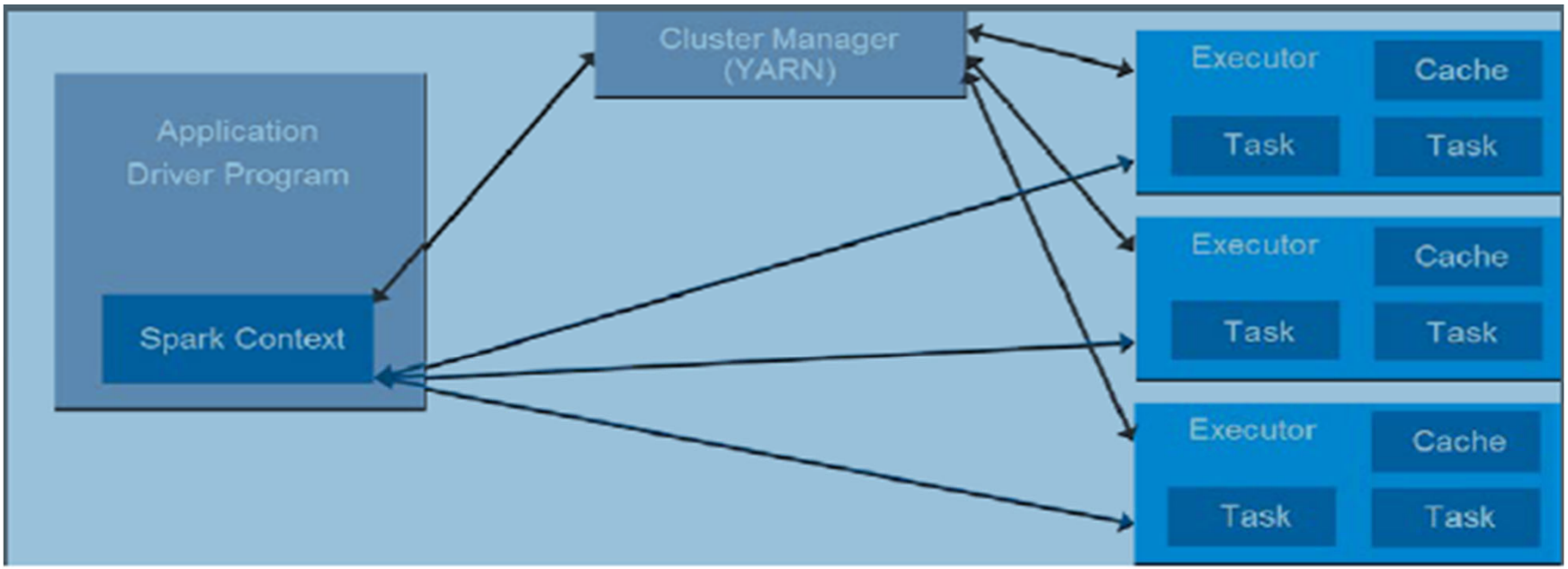


Requirements: Why Hadoop





Parallel processing: Task distribution



Number of executors: 16
Number of cores : 46
Batch size : 1024 , 512 , 2048 ..etc

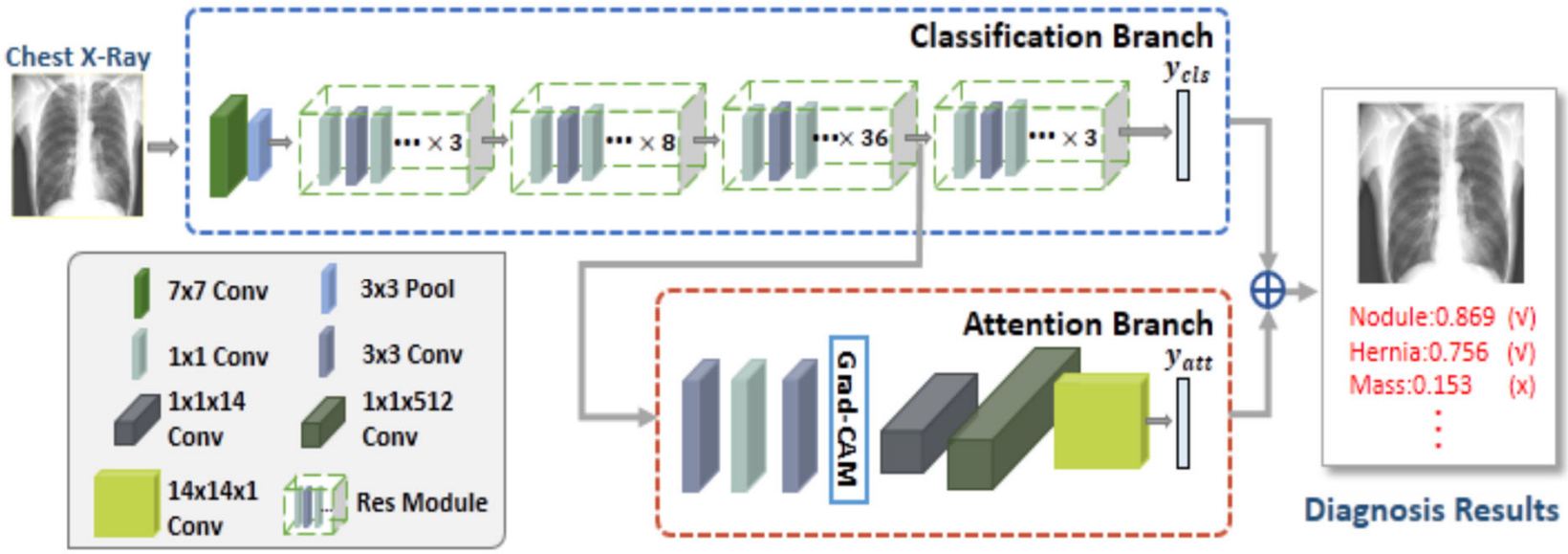
Deep learning using transfer learning

- Transfer learning is utilized to predict the diseases in images using
  - Resent
  - Densenet
  - VGG16
  - Inception

- The complete code is here :

  HTML_code_Chest_ray.html

Deep learning: workflow



Number of layers : 227 layers
Last layer : sigmoid

Our Results

TABLE II
OUR RESULTS IN COMPARSION WITH OTHER STATE OF THE ART RESULTS.

Class Name	OurAccuracy	Wang et al(2017)	Yao et al. (20
Atelectasis	0.794554080842	0.716	0.772
Cardiomegaly	0.873621830242	0.807	0.904
Effusion	0.867429446012	0.784	0.859
Infiltration	0.69914106984	0.609	0.695
Mass	0.815454190407	0.706	0.792
Nodule	0.720415340566	0.671	0.717
Pneumonia	0.722111574916	0.633	0.713
Pneumothorax	0.87348621061	0.806	0.841
Consolidation	0.784385714286	0.708	0.788
Edema	0.891658152995	0.835	0.882
Emphysema	0.884076076811	0.815	0.829
Fibrosis	0.78989952945	0.769	0.767
Pleural Thickening	0.759304768164	0.708	0.765
Hernia	0.740815766503	0.767	0.914
Average auc	0.80117	0.73814	0.80271