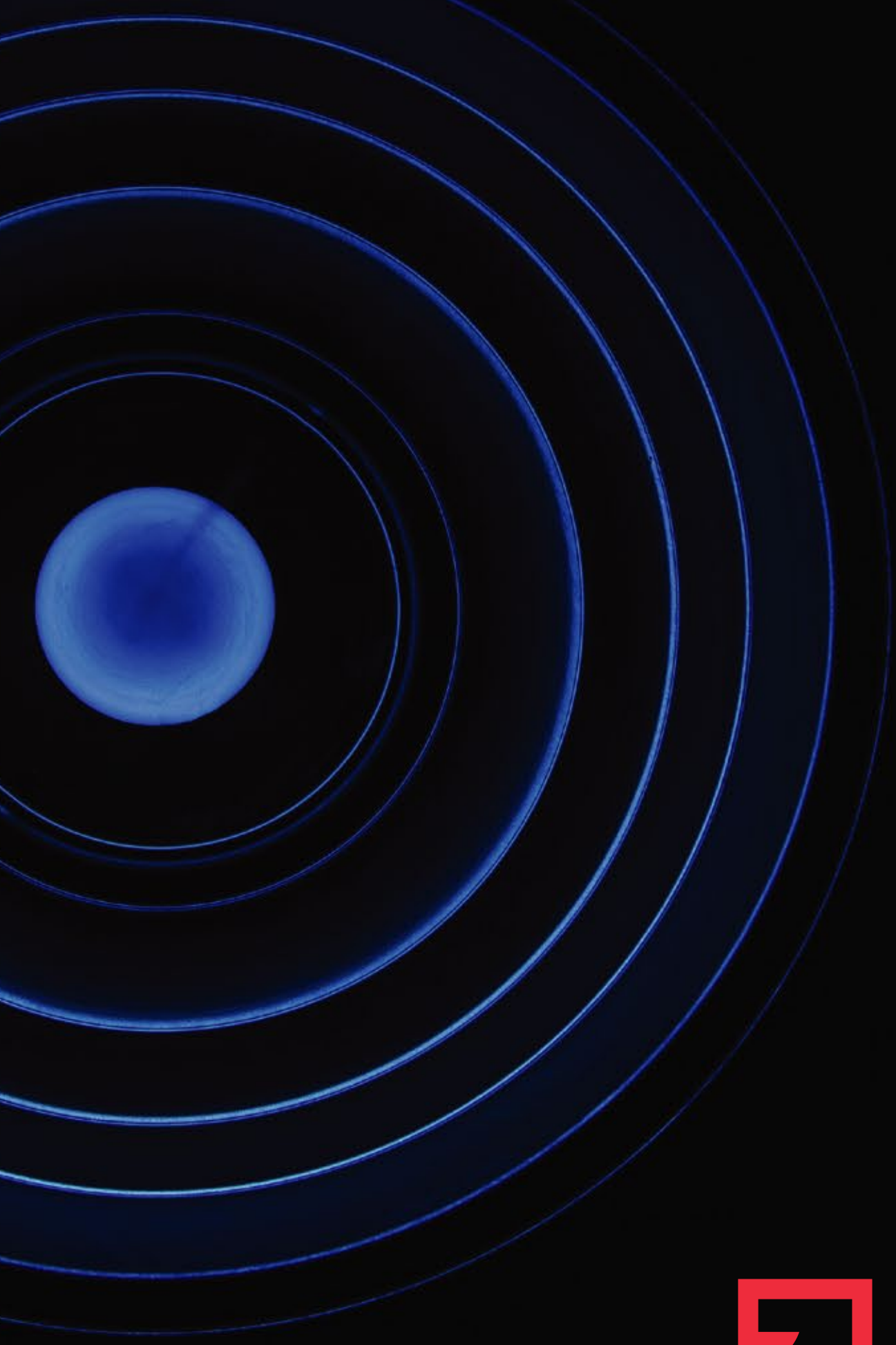
ISSUE 02
JANUARY 2018

DATA STREAK

A MONTHLY DIGEST ON ALL THINGS DATA



UpGrad

Founder's Speak

ML/AI and its Impact on Jobs

Growth in Machine Learning and Artificial Intelligence is one of the biggest shifts happening in the technology ecosystem right now. A major point being raised and debated is the impact of this progress on millions of jobs around the world. Closer home, there is a growing concern, and rightly so, that the IT industry which has fueled the growth of middle class in India, is under threat. With a lot of repetitive work being or set to be automated, this potentially puts the whole business model of the IT companies at risk.

While this is a very valid conversation, this is not the first time in history that an advancement in technology has lead to concerns over job losses. There was a massive concern in Europe that the Industrial Revolution was going to put thousands of weavers out of work. It actually spurred demand and ended up creating many more thousand manufacturing and machine operation jobs. The weaving job did not disappear. Nature of job changed and became less focussed on repetitive tasks which were taken over by the machines.

Historically, automation has always had a short term effect on jobs, but in the long term it has either changed the nature of the same job or actually created more jobs by increasing demand. As noted technologist Chris Dixon says, 'the set of human demands knows no bounds'. When the rapid pace of internet growth had everyone singing death warrants for multiple jobs, no one could predict the scale of new jobs like web and mobile dev, social media and digital marketing managers or even jobs like Uber drivers that would get created which would've been impossible without the internet. It's always easy to see which jobs will become redundant but it's almost impossible to know which new jobs will be created in the future because by definition, they don't exist yet.

No one knows exactly how this will play out, but I am cautiously optimistic that automation of repetitive tasks will increase human productivity and greatly and spur growth of higher skilled and more creative jobs. AI and automation will fundamentally change many industries, from manufacturing to healthcare, and these industries will need skilled people and partners to help them through this shift. So, as participants in this shift, what we as individuals and the IT companies in general can do best is focus on skills and job roles with strong tailwinds into the future. The nature of jobs and industries is changing rapidly and we have to stay a step ahead. By investing in continuously building ourselves for the jobs of tomorrow, we will be well positioned to take advantage of the vast benefits the rapid technology shifts have to offer.

Ravijot Chugh

Co-Founder, Product

UpGrad

ARCHITECTING BI SOLUTION

Student Article by Guru Charan Bulusu

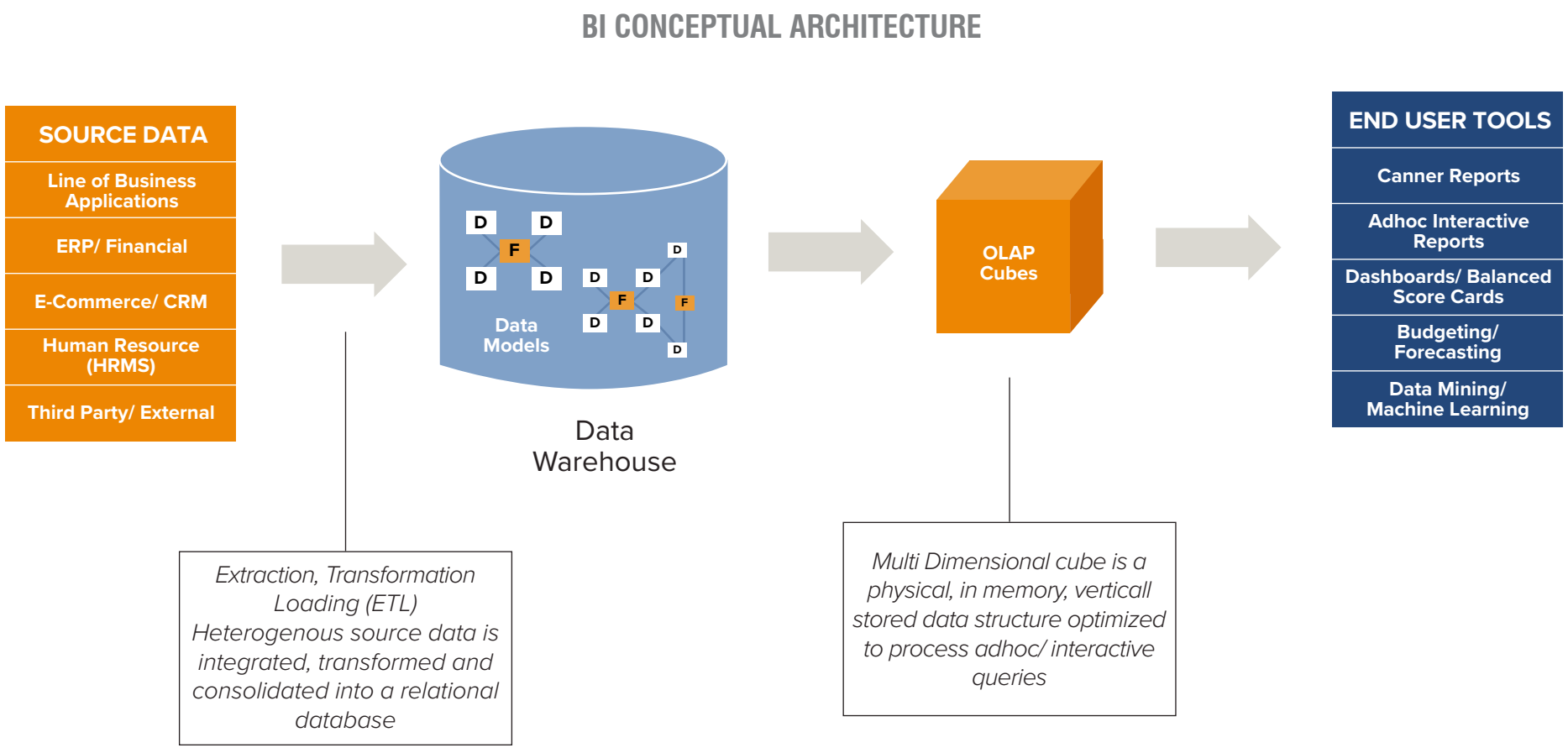
Software Architect, Microsoft

As a part of the discussion forum, there was a question from my peer as below. A very valid question considering options one must consider before building an "End to End BI solution".

How DW process works, how data feeds to the DW, how is batch job set-up to feed to DW, who exactly creates star schema. In one-line, end-to-end process till data stores in DW. Then how ET tool pulls data from DW. And the request was not to USE GOOGLE LINKS as answers.

Now, these questions cannot be answered in that forum, so I detailed it into a blog post. There are series of blog posts that cover from **Source Systems –> Building BI Systems –> Predicting Employee Attrition**. And all along, assets will be uploaded to GitHub if anyone wants to reuse. Also, adhering to general practice, let us understand BI solution from architectural, design and development standpoint to appreciate knowhow at each layer. This post details "Architecting BI Solution".

Below depicted is "Conceptual BI architecture" that is the most generic representation of BI solution on any platform and technology.



At an architectural level, below are questions that one needs to ask.

SOURCING DATA:

Regarding source data few of the questions are:

ACCESS METHODS:

High degree of variation in terms of how data can be extracted from source systems. Categorize application and identify different methodologies to pull data.

- **Homegrown application, generally extraction at data layer is possible.**
- **ERP software like SAP, there is NO way to extract data from DB layer. Instead, such application provides API and other access methodologies to pull data. One need to use only those to pull data.**
- **Cloud-based SAAS solution only has API methods to pull data.**

Generally, extraction at application/middle tier layer is the ONLY option for Third Party/External applications.

DATA FORMATS:

Different applications provide different data formats. If accessing DB directly, it becomes straightforward but generally one ends up using different access methods as aforementioned. When extracting data from API layer from source systems variations in data format will arise.

- **Any RDBMS structured data**
- **API Layer extraction**
- **Different file formats (CSV / TSV).**
- **JSON files**

DATA VOLUMES:

Next aspect architect needs to be aware, is the volume of delta/differential data.

- **Does source system provide methods to pull delta data?**
- **If so, data volume (in terms of GBs/TBs) of differential daily extraction.**

Other points to consider would be, Network Throughput, the impact of extraction on the performance of source systems, Security aspects (authorization/auditing), allowable staleness of data in DWH (D – 1).

Once the architect has a clear picture of various sources and related inputs, next question one has to answer is, should data be hosted in an ODS (Operational Data Store) or can it be directly loaded into dimensional models (Start/Snowflake Schema).

REQUIREMENT FOR OPERATIONAL DATA STORES:

Operational data stores retain the same schema as source systems and there is no schema level impedance between Source System (Relational Models) and ODS Schema. The only difference is, ODS holds data from all source systems for a limited period of them before loading into models of a data warehouse.

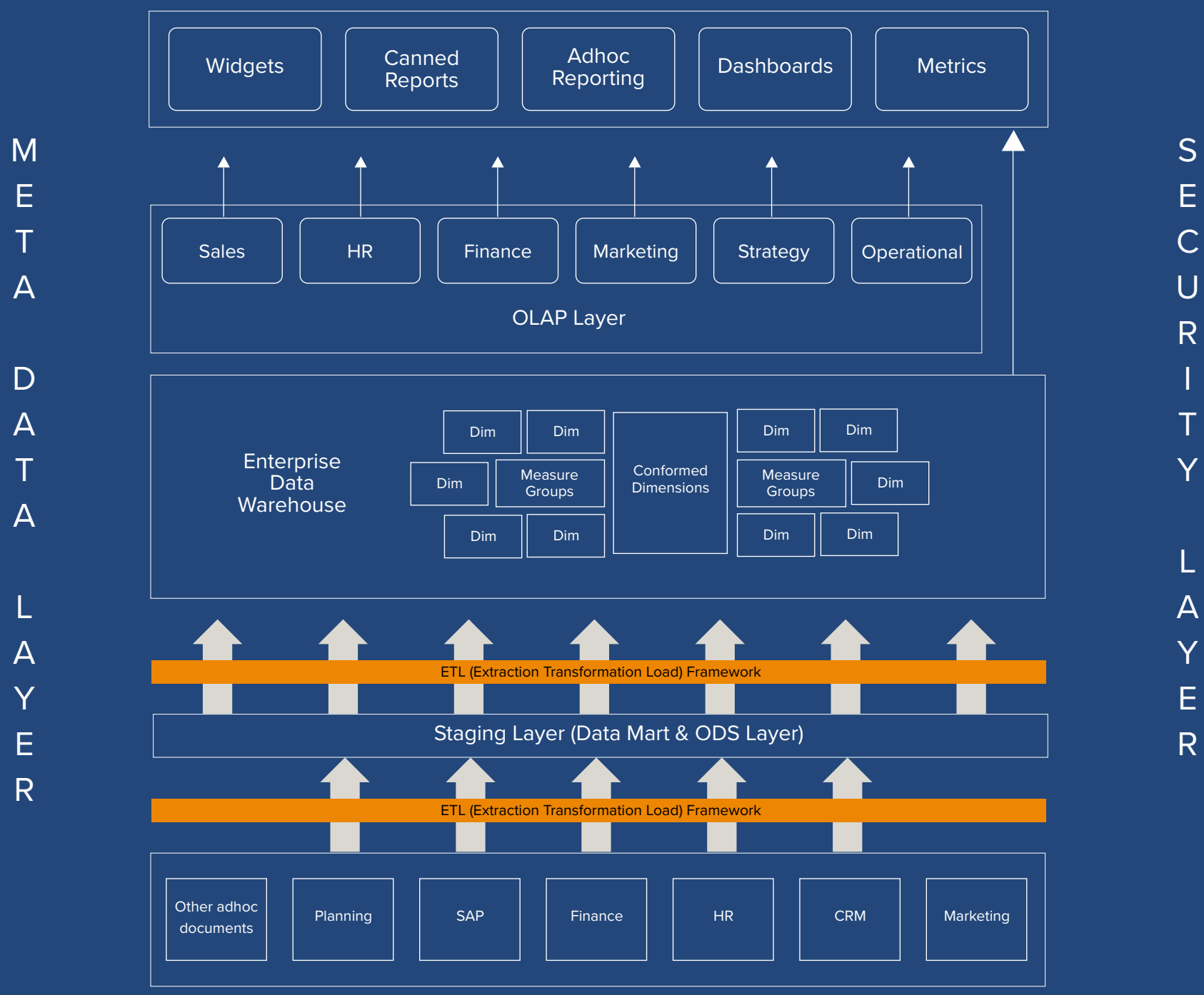
- **Source System: Normalized to store data model for that specific application.**
- **ODS: Normalized data models, hosted on all source systems.**
- **DWH: De-normalized data models (as designed by data models).**

To understand if ODS is required in a BI solution, these questions need to be answered.

- **Does current OLTP system support reporting or are systems too stressed (in terms of resources) that running reports will impact business transactions.**
- **For running reports, does data need to be integrated with other transaction systems? Like for example, to understand a customer journey end to end on a site, not just Order System but Web Log transaction systems need to integrate to understand customer journey. Similar is the case of taking inventory stock across various stores.**

But be careful about ODS. Not many people like it and maybe because “Data Mart”, “Data Warehouse”, “ODS”, “Report Data Store” are used interchangeably. But if ODS is built, it becomes the source for Data Warehouse.

ENTERPRISE DATA WAREHOUSE DATA FLOW DIAGRAM



ETL:

Data extraction, transformation and load enable to move data from Source Systems to either ODS (if built) or to data models in Data Warehouse. A generic architecture data flow architecture for a BI solution depicted below. Notice those highlighted in yellow indicating ETL layer.

General capability/architectural questions for ETL listed, but not limited to below

- **Support for diverse data sources (from Data Repositories (SQL / NoSQL) to ERP (SAP/PeopleSoft), Web Resources (Web Services, RSS Feeds).**
- **Availability of high-performance provider both for source and destination systems. (This is going to be key for performance of ETL layer).**
- **ETL Scalability is the key. Generally, ETL systems due to the in-memory continuous pipe are good for row-based transformation but not good for Set-based operations. For example, if a column of a row (row based operations) needs to be transformed, ETL is the best tool, but if data from multiple tables need to join, aggregated (set based operations) database technologies are better.**
- **Also in a complex ETL workflow, there may be requirement for integration with messaging middleware like MQ Series or MSMQ or Biz Talk. Such requirement if any, needs to be gathered. For example, SAP data may need to be extracted using middleware like Biz Talk and then ETL process will initiate.**

Finally, requirements for Auditing, error handling & logging, adherence to compliances are going to be key points.

DWH:

Data Warehouse at an architecture level is more aligned towards designing dimension models for required subject areas and adhered principles. Other non-functional requirements like Size/volume of data etc come into play. Below depicted is a generic HR model that would help capture employee (Active/Left) information that could be later used for predicting employee attrition.

When moving to physical/deployment architectures, DBA skills will take a long way in implementing large yet scalable and highly performant databases that host dimensional models. Also, as a general practice and recommendation by Ralph Kimball (Father of the dimensional model), relationships between dimensions and facts are captured in a “BUS MATRIX”.

DIMENSION(S)	HEAD COUNT	QUIT NUMBERS
Employee	X	X
Qualification	X	X
Role	X	X
Department	X	X
Project	X	X
Experience Level	X	X
Salary Level	X	X
Date (Role Palying)	X	X
Leaving Reason		X

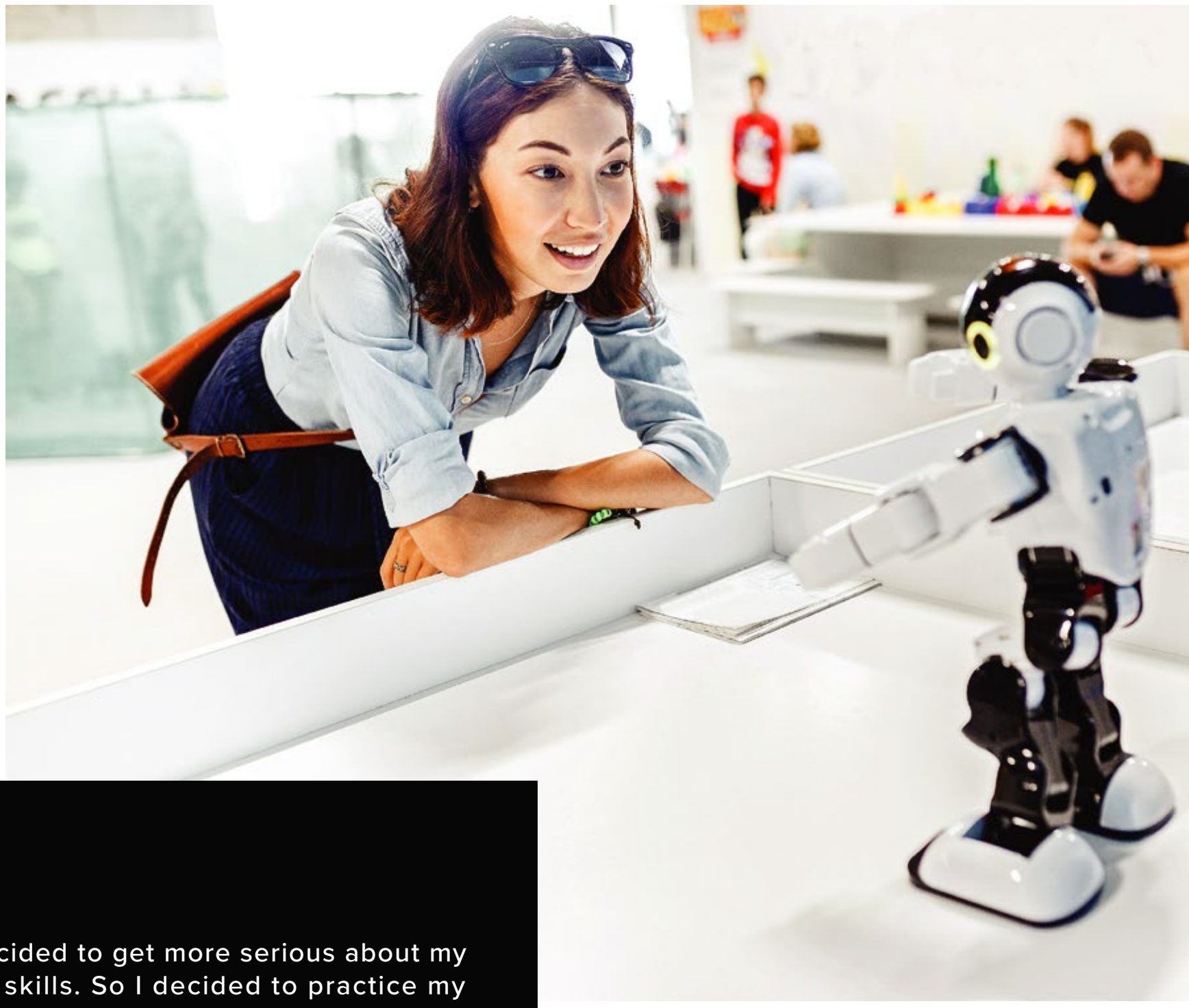
There are two other layers of architecture above, OLAP and Reporting Layer that will be covered next in my blog.

Guru Charan’s Blog Link

https://abhyast.wordpress.com/2017/04/21/architecting-bi-solution

Machine Learning Contests on Kaggle

MY INTRODUCTION & EXPERIENCES



Recently I decided to get more serious about my data science skills. So I decided to practice my skills, which led me to Kaggle. The experience has been very positive.

When I arrived at Kaggle, I was confused about what to do and how everything works. This article will help you overcome the confusion that I experienced.

I joined the Redefining Cancer Treatment contest, because it was for a noble cause. Also, the data was more manageable because it was text based.

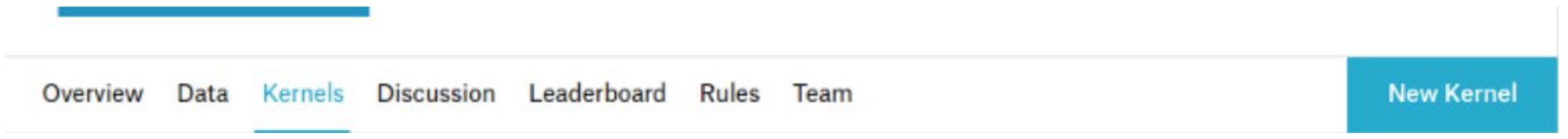
WHERE TO CODE

What makes Kaggle great is that you don't need a cloud server that creates results for you. Kaggle has a feature where you can run scripts and notebooks inside Kaggle for free, as long as they finish executing within an hour. I used Kaggle's notebooks for many of my submissions, and experimented with many variables. Overall it was a great experience.

PARMINDER SINGH

Writer, Developer and Data Scientist

https://trion.me



That New Kernel button is your friend!

For the contests, you need to use images or have a large corpus of text. And you will need a fast personal computer (PC) or a cloud container. My PC is crappy, so I used Amazon Web Services' (AWS) c4.2xlarge instance. It was powerful enough for the text and costed only $0.40 per hour. I also had a free $150 credit from the GitHub student developer pack, so I didn't need to worry about the cost.

Later when I took part in the Dog Breed Identification playground contest, I worked a lot with images, so I had to upgrade my instance to g2.2xlarge. It costed $0.65 per hour, but it had graphics processing unit (GPU) power, so that it could compute thousands of images in just a few minutes.

The instance g2.2xlarge was still not large enough to hold all of the data I worked with, so I cached the intermediate data as files and deleted the data from RAM. I did this by using del <variable name> to avoid ResourceExhaustionError or MemoryError . Both were equally disheartening.

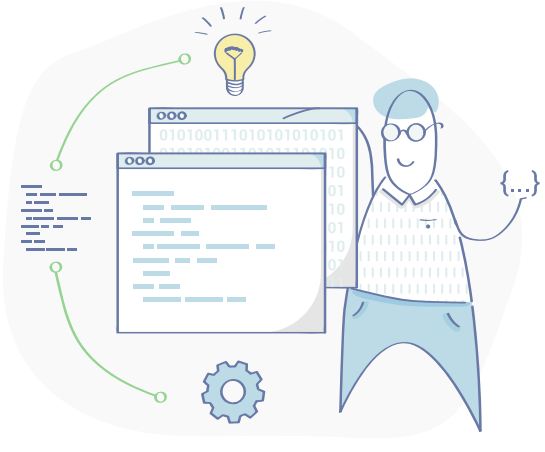
HOW TO GET STARTED WITH KAGGLE COMPETITIONS

It's not as scary as it sounds. The Discussion and Kernel tabs for every contest are a marvellous way to get started. A few days after the start of a contest, you will see several starter kernels appear in the Kernel tab. You can use these to get started.

Instead of handling the loading and creation of submissions, just deal with the manipulation of data. I prefer the XGBoost starter kernels. Their codes are always short and are ranked high on leaderboards.

Extreme Gradient Boosting (XGBoost) is based on the decision tree model. It is very fast and amazingly accurate, even on default variables. For large data I prefer to use Light Gradient Boosting Machine (LightGBM). It is similar in concept to the XGBoost, but approaches the problem a bit differently. There is a catch, it is not as accurate. So you can experiment using LightGBM, and when you know it is working great, switch to XGBoost (they have a similar API).

Check the discussions every few days to see if someone has found a new approach. If someone does, use it in your script and test to see if you benefit from it.



HOW TO GO UP IN THE LEADERBOARD

So you have your starter code cooked and want to rise higher? There are many possible approaches:

- **Cross validation (CV): Always split the training data into 80% and 20%. That way when you train on 80% of the data, you can manually cross-check with 20% of the data to see if you have a good model. To quote the discussion board on Kaggle, "Always trust your CV more than the leaderboard." The leaderboard has 50% to 70% of actual test set, so you cannot be sure about the quality of your solution based on the percentages. Sometimes your model might be great overall, but bad on the data, specifically in the public test set.**
- **Cache your intermediate data: You will do less work next time by doing this. Focus on a specific step rather than running everything from the start. Almost all python objects can be pickled, but for efficiency, always use .save() and .load() functions of the library you are using for your code.**
- **Use GridSearchCV: It is a great module that allows you to provide a set of variable values. It will try all possible combinations until it finds the optimal set of values. This is a great automation for optimization. A finely tuned XGBoost can beat a generic neural network in many problems.**
- **Use the model appropriate to the problem: Using a knife in a gunfight is not a good idea. I have a simple approach: For text data, use XGBoost or Keras LSTM. For image data, use Pre-trained Keras model (I use Inception most of the time) with some custom bottleneck layers.**
- **Combine models: Using a kitchen knife for everything is not enough. You need a Swiss army knife. Try combining various models to get even more accurate information. For example, Inception plus the Xception model work great for image data. Combined models take a lot of RAM, which g2.2xlarge might not provide. So avoid them unless you really want to get that accuracy boost.**
- **Feature extraction: Make the work easier for the model by extracting multiple simpler features from one feature, or combining several features into one feature. For example, you can extract the country and area code from a phone number. Models are not very intelligent, they are just algorithms that fit data. So make sure that the data is appropriate for optimal fit.**

WHAT ELSE TO DO ON KAGGLE

Other than being a competition platform for data science, Kaggle is also a platform for exploring datasets and creating kernels that explore insights into the data.

So you can choose any dataset out of the top five that appear on the datasets page, and just go with it. The data might be weird, and you might experience difficulty as a beginner. What matters is that you analyze data and make visualizations related to it, which contributes to your learning.

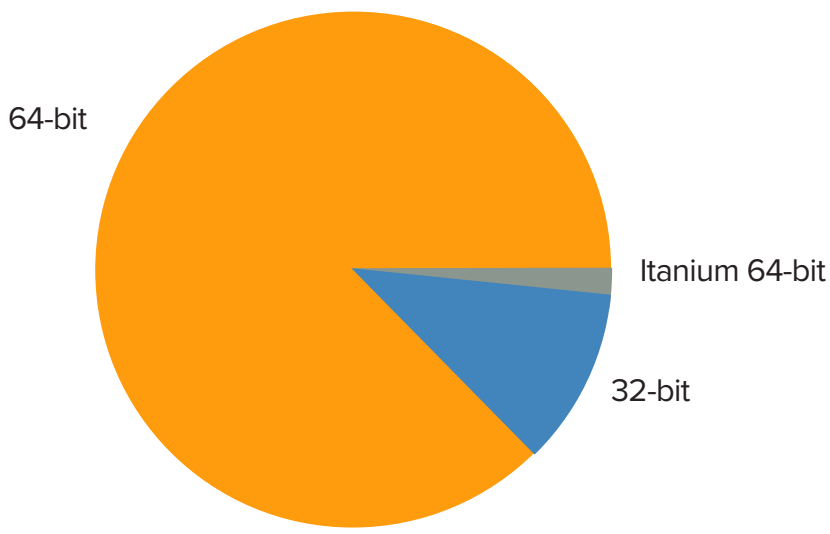
WHICH LIBRARIES TO USE FOR ANALYSIS

For ***visualizations, explore seaborn and matplotlib libraries***
For ***data manipulation, explore NumPy and pandas***
For ***data preprocessing, explore sklearn.preprocessing module***

Pandas' library has some basic plot functions too, and they are extremely convenient.

```
intel_sorted["Instruction_Set"].value_counts().plot(kind='pie')
```

The one-line code above made a pie chart with "Instruction_Set." And the best thing is that it still looks pretty.

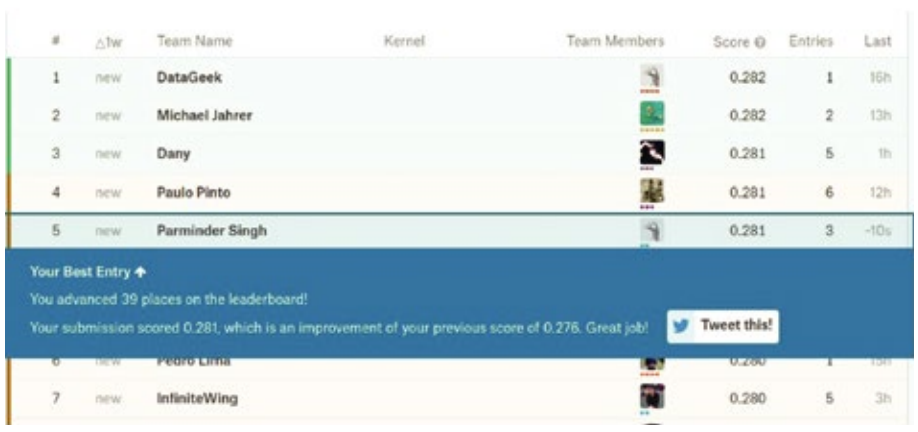


WHY DO ALL THIS?

Machine learning is a beautiful field with lots of on-going development. Participating in these contests will help you learn a lot about algorithms and the various approaches to data. I myself learned a lot of these things from Kaggle.

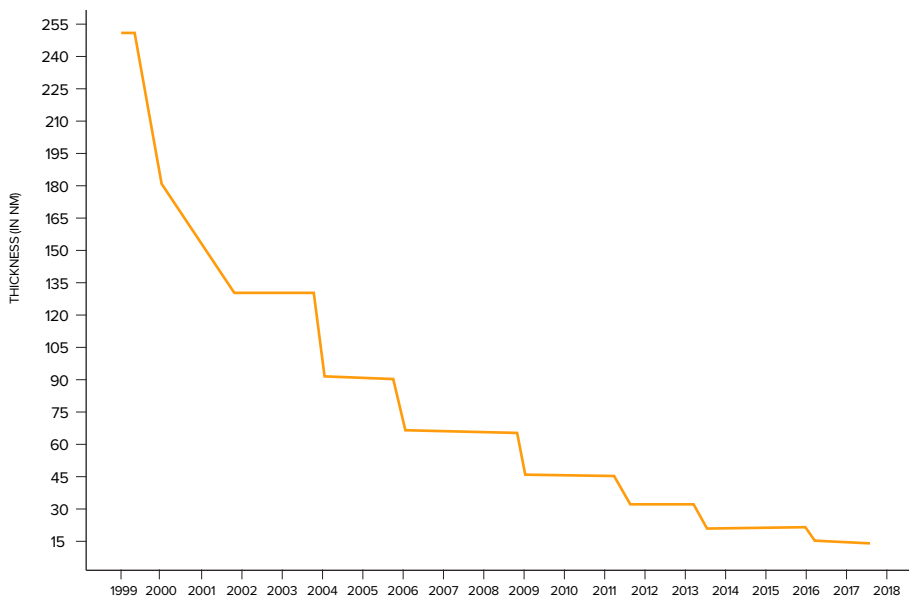
Also, to be able to say, **"My AI is in the top 15% for <insert contest name here>"** is pretty dope.

SOME EXTRAS FROM MY JOURNEY



IN TOP 5

The graph below represents my kernel's exploration of the Intel CPU dataset on Kaggle:



My solution for the Redefining Cancer Treatment contest:

RANKED 217

That's all folks.

Thanks for reading. I hope I made you feel more confident about participating in Kaggle's contests.

See you on the leaderboards.

Parminder Singh's Blog Link

https://medium.freecodecamp.org/what-i-learned-from-kaggle-contests-d3123e17a36b

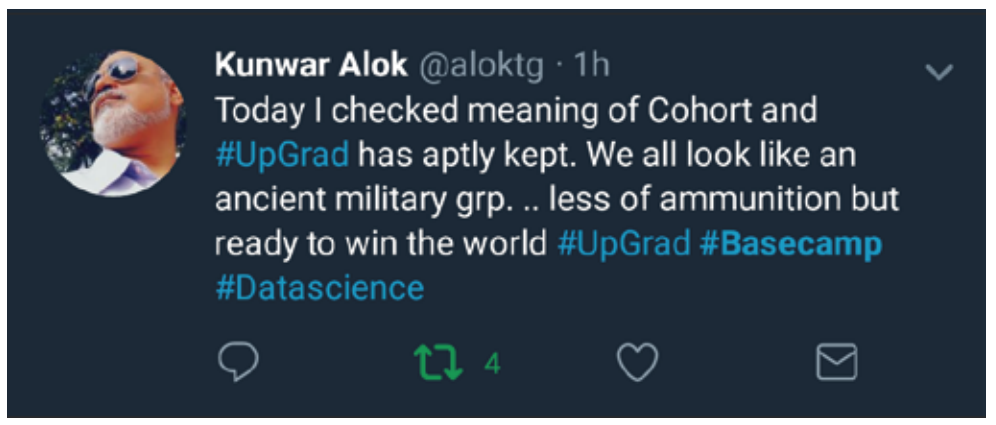
YouTube Talks to check out

https://www.youtube.com/watch?v=aircAruvnKk
https://www.youtube.com/watch?v=b8g-8T0amuk

We successfully completed the 1st phase of BaseCamp for Cohorts 3 & 4 in the months of November and December across 4 cities - Bangalore, Mumbai, Hyderabad and Delhi. BaseCamp provided us with a great platform to meet our learners and it was amazing to see such huge turnouts across all cities. More than 240 learners participated in the event. The highlights of this BaseCamp were sessions on Kaggle conducted by Mr Vikram Jha, CEO of Pucho and Bishwarup Bhattacharjee Kaggle grandmaster and Data Scientist at Aditya Birla group who covered “how taking up projects and case studies on kaggle will enhance your skills and will give you an edge over others. Another highlight of the event was the Career Talk by our In-house Experts Mr Jayadev Mahalingam and Varun Singh Director-Strategic Alliances and Outreach at UpGrad. They discussed the key skills that recruiters look for and action points that help one successfully transition to a career of their choice.

Learners went away with a lot of learning along with some cool prizes, UpGrad souvenirs and happy memories. It was great to see you guys bonding well with your cohort mates.

Don't fret if you missed this one, we are coming up with many more BaseCamps in this New Year. We will keep you posted.



UP-SKILLED IN DATA SCIENCE

WITH IIIT BANGALORE & UPGRAD

An UpGrad Student Testimonial

ABHINAY BANDARU

PRODUCT ANALYST

HEAD infotech

I am a Data Analyst at Head Infotech Pvt Ltd, a gaming company in Hyderabad. I have done my B.Tech from NIT Surat in Electronics & Communication Engineering and it was clear in my final year of graduation that I was not going to continue my career as an Electronics Engineer due to various reasons. This is the point where I thought of a career transition. I researched about the areas where I can do well and since I was good at Mathematics, Statistics and Physics right from my childhood, it helped me in taking the decision to shift towards data science. It is a field where you can play with mathematics and logic to help answer difficult business problems using analytics.

I got to know about UpGrad's PGDM in Data Science, with a certification from IIIT Bangalore, through one of my friends. After checking the course content, I found the pedagogy of content to be well organized. The reviews were very good as it was taught by professors from IIIT Bangalore. Even after that, I had a lot of discussions with many people along with the UpGrad mentor before arriving at the decision to take the course, as it was priced a little high for me. Finally, I enrolled in the course and it was the best decision I took just after completing my B.Tech.

I had taken this online course leaving all the electronics engineering knowledge and completely concentrating on Data Science. The content taught by IIIT-B professors made me more excited about this field. The support of UpGrad student mentors in the course was a crucial part. I followed them and they helped a lot in clarifying my queries. Placement assistance was provided where you are perfectly guided on how to start a career as a Data Analyst. When everything was going fine while learning all these, I got a call from Head Infotech Pvt. Ltd. to attend for an interview for Data Analyst role. After several rounds, finally, all my efforts after the graduation from my home, paid off - achieving career start as a Data Analyst. I would like to thank IIIT-B professors and UpGrad for providing course content so systematically, and also Head Infotech Pvt. Ltd. for believing in me to join their family.

At some point in your life, you have to take a calculated risk to achieve success in life and this was it.

More Student Testimonials

https://www.linkedin.com/pulse/professional-career-personal-development-upgrad-data-anshu-srivastava/

https://www.linkedin.com/pulse/my-upgrad-story-vivek-mohan-singh

Recent Career Transitions

Gorenti Vinay
Cohort 2

FRESHER

ANALYTICS ASSOCIATE

159 SOLUTIONS

Mukund Kumar
Cohort 2

PROJECT ENGINEER

TECOSIM
better life by simulation

CONSULTANT
(FPM & PREDICTIVE ANALYTICS)

4TECHNOLOGIES
BUSINESS FIRST

Snigdha Prakash
Cohort 2

SENIOR SOFTWARE ENGINEER

edgeverve
An Infosys company

DATA SCIENTIST

TATA
TATA CONSULTANCY SERVICES

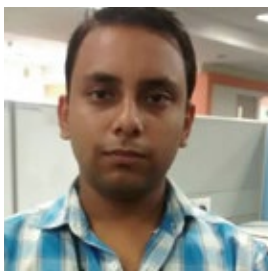
Nirmal Maheshwari
Cohort 2

SOFTWARE ENGINEER

EKO/-

ANALYST/ POWER BI

Q3 technologies
Building Quality into Software



Soumadiptya Chakraborty
Cohort 2

PROGRAMMER ANALYST

Cognizant

DATA SCIENTIST

Cognizant

Recent Career Transitions

Kusuma Rajesh
Cohort 2

FRESHER

DATA SCIENTIST

COPPIUS TECH

Manoj Gondimalla
Cohort 2

PRINCIPAL ANALYST

citi

SPECIALIST ROLE

Fidelity INVESTMENTS

Harsha Raikar
Cohort 3

TECHNICAL ARCHITECT (.NET)

ISG INFORMATION SERVICES GROUP

SOLUTIONS ARCHITECT (MACHINE LEARNING)

Huron CONSULTING GROUP

Jai Singh Bhagat
Cohort 3

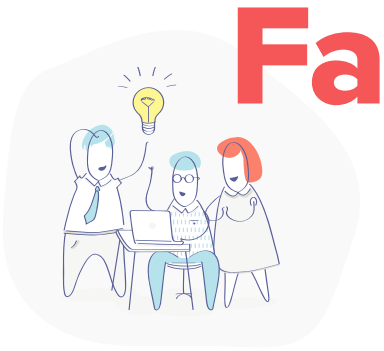
ASSISTANT SYSTEMS ENGINEER

TATA CONSULTANCY SERVICES

REPORT DEVELOPMENT & ANALYTICS (BIRST)

TATA CONSULTANCY SERVICES

Fun Facts



90% of the entire world's data was created in the last couple of years. While billions are being spent on making sense out of these available data, less than **0.5%** are getting analysed and made use of. Such mind-boggling extent of it is bound to give out some mind-startling results!

While classifying data, Tata Consultancy Services Limited (TCS) has looked at how much of companies' data was structured versus unstructured, as well as how much was generated internally versus externally. It found that **51%** of data is structured, **27%** of data is unstructured and **21%** of data is semi-structured. A much higher than anticipated percentage of data was not structured — either unstructured or semi-structured and a little less than a quarter of the data was external.

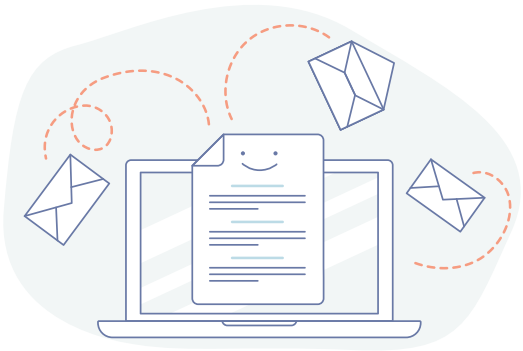
Quiz Time

Crossing the Bridge Puzzle

Four people need to cross a rickety bridge at night. Unfortunately, they have only one torch and the bridge is too dangerous to cross without one. The bridge is only strong enough to support two people at a time. Not all people take the same time to cross the bridge. Times for each person: 1 min, 2 mins, 7 mins and 10 mins. What is the shortest time needed for all four of them to cross the bridge?

10 Coins Puzzle

You are blindfolded and 10 coins are place in front of you on table. You are allowed to touch the coins, but can't tell which way up they are by feel. You are told that there are 5 coins head up, and 5 coins tails up but not which ones are which. How do you make two piles of coins each with the same number of heads up? You can flip the coins any number of times.



3,247 billion e-mail messages are sent each day. Up to **80%** of them are spam.

To share your stories/articles/blogs, write to us at
pgdds@upgrad.com

FIND US HERE

UpGrad Education Private Limited,
Nishuvi, 75, Dr. Annie Besant Road,
Worli, Mumbai – 400018.