Predictive Analytics & Machine Learning

UNIT 1:

Issues with big data


Big Data

Challenge #1: Insufficient understanding & acceptance of big data

Oftentimes, companies FAIL to know even the basics: what big data actually is, what its benefits are, what infrastructure is needed, etc.

Without a clear understanding, a big data adoption project risks to be doomed to failure.

Companies may waste lots of time and resources on things they don't even know how to use.

If employees don't understand big data's value and/or don't want to change the existing processes for the sake of its adoption, they can resist it and impede the company's progress.

Solution:

Big data, being a huge change for a company, should be accepted by TOP management 1st ^ then down the ladder.

To ensure big data understanding & acceptance at all levels, IT departments need to organize numerous trainings & workshops.

The implementation & use of the new big data solution need to be monitored and controlled.

However, top management should not overdo with control because it may have an adverse effect.

Challenge #2: Confusing variety of big data technologies

- Easy to get LOST in the variety of big data technologies now available on the market.
- Do you need Spark or would the speeds of Hadoop MapReduce be enough?
- Is it better to store data in Cassandra or HBase?
- Finding the answers can be tricky.
- Even easier to choose poorly, if you are exploring the ocean of technological opportunities without a clear view of what you need.

Solution:

If you are new to the world of big data, trying to **seek professional** help would be the right way to go.

You could hire an **expert** or turn to a **vendor** for big data consulting.

In both cases, with joint efforts, you'll be able to work out a strategy and, based on that, choose the needed technology stack.

Challenge #3: Paying loads of money

- Big data adoption projects entail lots of expenses.
- If you opt for an ***on-premises*** solution, you'll have to mind the costs of new ***hardware***, new ***hires*** (administrators & developers), ***electricity***, etc.
- Additionally, you'll need to pay for the development, setup, configuration & maintenance of new ***software***.
- If you decide on a cloud-based big data solution, you'll still need to ***hire staff*** and pay for ***cloud services***, big data solution ***development*** as well as setup & maintenance of needed ***frameworks***.
- Moreover, in both cases, you'll need to allow for future ***expansions*** to avoid big data growth getting out of hand and costing you a fortune.

Solution:

The particular salvation of your company's wallet will depend on your company's specific technological needs & business goals.

For instance, companies who want flexibility benefit from cloud. While companies with extremely harsh security requirements go on-premises.

There are also hybrid solutions when parts of data are stored & processed in cloud and parts – on-premises, which can also be cost-effective.

Resorting to data lakes or algorithm optimizations (if done properly) can also save money:

- ***Data lakes*** can provide cheap storage opportunities for the data you don't need to analyze at the moment.
- ***Optimized algorithms***, in their turn, can reduce computing power consumption by 5 to 100 times. Or even more.

All in all, the key to solving this challenge is properly analyzing your needs and choosing a corresponding course of action.

Challenge #4: Complexity of managing data quality

Data from diverse sources

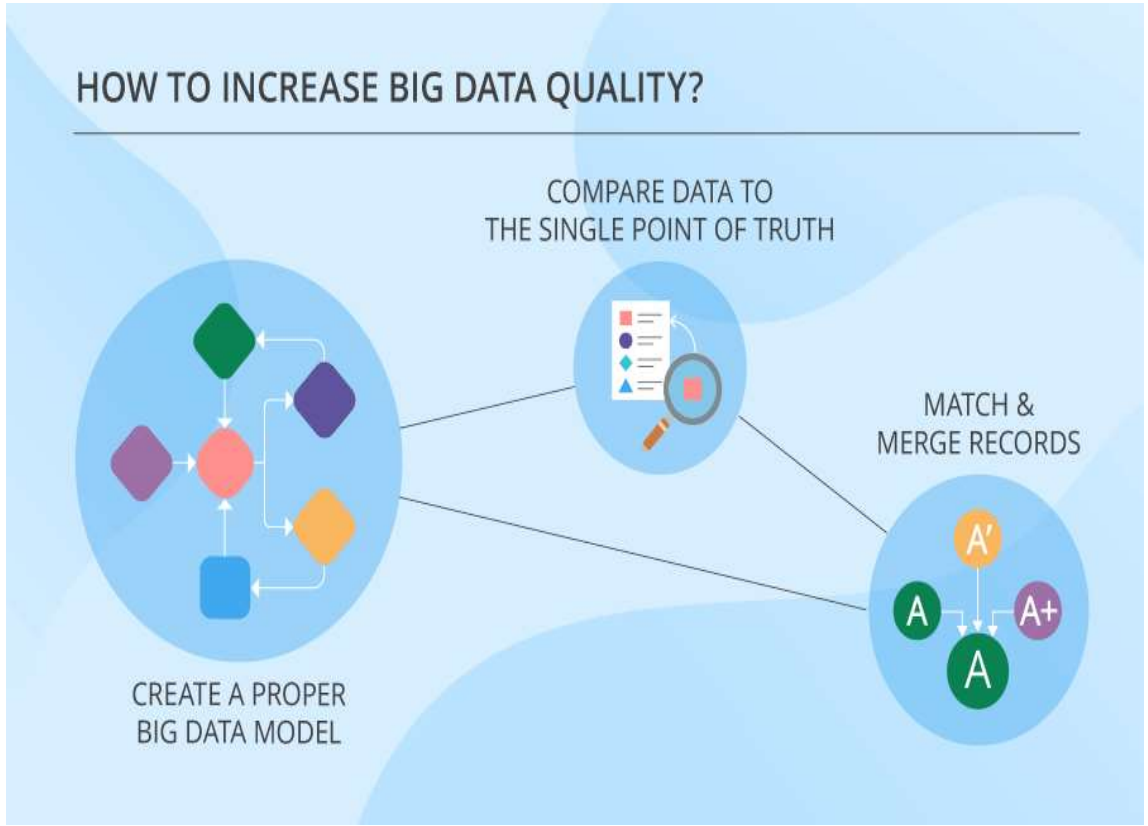
- Sooner or later, you'll run into the problem of *DATA INTEGRATION*, since the data you need to analyze comes from *DIVERSE SOURCES* in a variety of different formats.
- For instance, ecommerce companies need to analyze data from website logs, call-centers, competitors' website 'scans' and social media.
- Data *formats* will obviously differ, and matching them can be problematic.
- For example, your solution has to know that skis named *SALOMON QST 92 17/18*, *Salomon QST 92 2017-18* and *Salomon QST 92 Skis 2018* are the same thing, while companies **IAQS** and **laQs** are not.

Challenge #4: Complexity of managing data quality

Unreliable data

- Nobody is hiding the fact that big data isn't 100% accurate
- Not only can it contain ***WRONG*** information, but also ***DUPLICATE*** itself, as well as contain CONTRADICTIONS.
- And it's unlikely that data of extremely ***INFERIOR*** quality can bring any useful insights or shiny opportunities to your precision-demanding business tasks.

Solution:



- There is a whole **BUNCH** of techniques dedicated to cleansing data.
- Your big data needs to have a proper MODEL. Only after creating that, you can go ahead & do other things, like:
  - ***COMPARE*** data to the single point of truth (for instance, compare variants of addresses to their spellings in the postal system database).
  - Match & merge records, if they relate to the same entity.
- But mind that big data is never 100% accurate. You have to know it & deal with it.

Challenge #5: Dangerous big data security holes

Security challenges of big data are quite a VAST issue.

Quite often, big data adoption projects put security off till LATER stages.

Big data technologies do EVOLVE, but their security features are still NEGLECTED, since it’s hoped that security will be granted on the application level.

And what do we get? Both times (with technology advancement & project implementation) big data security just gets CAST ASIDE.

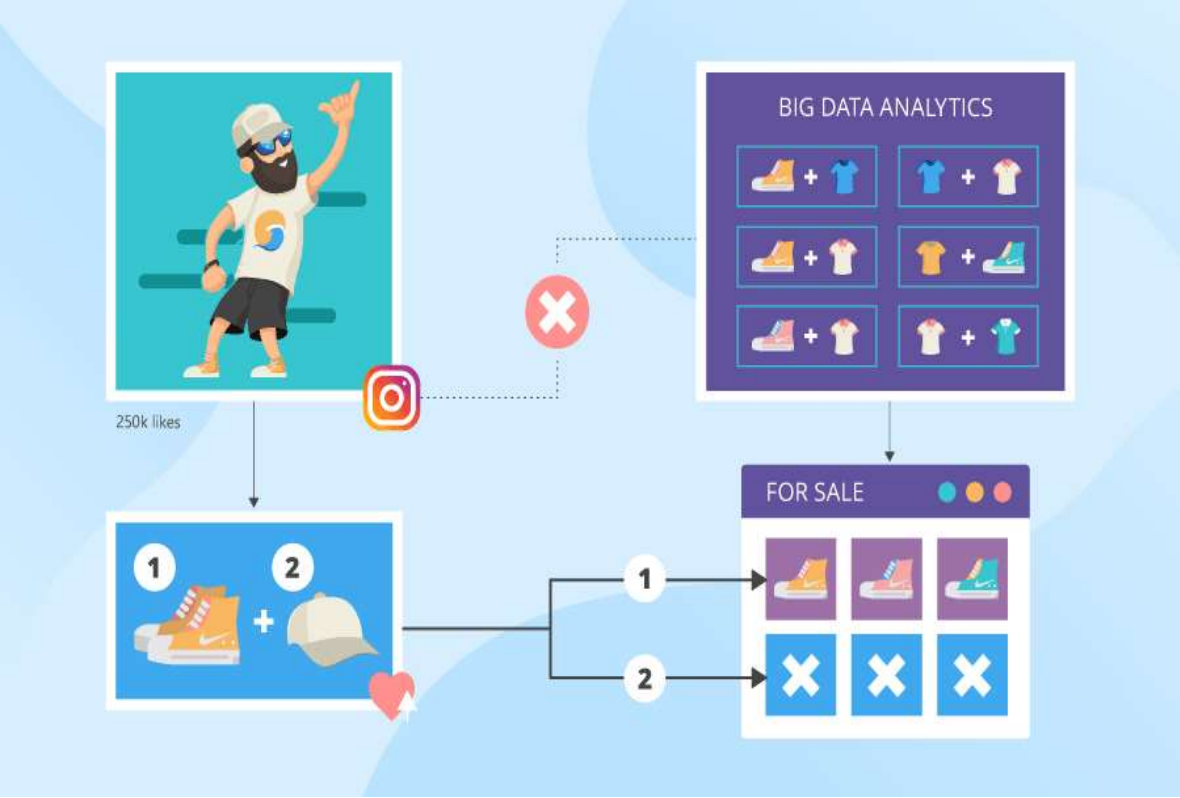
Solution:

The precaution against your possible big data security challenges is putting security FIRST.

It is particularly important at the stage of DESIGNING your solution’s architecture.

Because if you don’t get along with big data security from the VERY START, it’ll bite you when you least expect it.

Challenge #6: Tricky process of converting big data into valuable insights



- *An example*: your super-cool big data analytics looks at what item pairs people buy (say, a needle & thread) solely based on your historical data about customer behavior.
- Meanwhile, on Instagram, a certain soccer player posts his new look, and the 2 characteristic things he's wearing are white Nike sneakers & a beige cap.
- He looks good in them, and people who see that want to look this way too.
- Thus, they rush to buy a similar pair of sneakers & a similar cap.
- But in your store, you have only the sneakers. As a result, you lose revenue & maybe some loyal customers.

Solution:

The reason that you failed to have the needed items in stock is that your big data ***tool DOESN'T ANALYSE data from social networks*** or competitor's web stores.

While your rival's big data among other things does note trends in social media in near-real time. And their shop has both items and even offers a 15% discount if you buy both.

The idea here is that you need to create ***a PROPER SYSTEM*** of factors and data sources, whose analysis will bring the needed insights, and ensure that nothing falls out of scope.

Such a system should often ***include EXTERNAL SOURCES***, even if it may be difficult to obtain and analyze external data.

Challenge #7: Troubles of upscaling

The most typical feature of big data is its dramatic **ability to GROW**.

And one of the **MOST SERIOUS** challenges of big data is associated exactly with this.

Your solution's design may be thought through & adjusted to upscaling with no extra efforts.

But the real problem isn't the actual process of introducing new processing & storing capacities.

It lies in the COMPLEXITY of scaling up so, that your system's performance doesn't decline and you stay within budget.

Solution:

- The 1st & foremost precaution for challenges like this is a ***decent architecture*** of your big data solution.
- Another highly important thing to do is **designing your big data *algorithms*** while keeping future upscaling in mind.
- But besides that, you also ***need to plan*** for your system's maintenance and support so that any changes related to data growth are properly attended to.
- And on top of that, holding ***systematic performance audits*** can help identify weak spots and timely address them.

Others.....

- High Cost of Data Solutions
- Low Quality and Inaccurate Data
- Compliance Hurdles
- Using Data for Meaning
- Accessibility
- Pace of Technology
- Lack of Skilled Workers
- Processing Large Data Sets