NASSCOM® Engineering R&D

In Association with

SAP | NASSCOM® Center of Excellence-IoT & AI A MeitY Initiative with Govt. of Karnataka, Haryana, Gujarat & AP | futureskills

Questions Asked	Answers Given
What kind of Big Data Architecture is Lambda/Kappa?	Both Lambda and Kappa architectures can be used in building big data and depending on use case scenarios one can be chosen over the other, for example if you need lot of window, complex algorithms, batch then Lambda would suit well, but if it is simple real time then
What are the major challenges foreseen in the industries that would have been implemented in Big data analytics over the upcoming years? Any specific industries where the challenges are more? If so, why? And how can we solve the problem today having forseen it?	I would say each industry has set of challenges that are being currectly addressed, please see some articles that provide some insights (https://www.forbes.com/sites/louiscolumbus/2018/05/23/10-charts-that-will-change-your-perspective-of-big-datas-growth/#b8b77da29268)
How did you handle Data Enrichment on Streaming Data?	There are multiple ways to enrich data for example one could use apache flink or spark to perform actions like reference data lookup, pre-loading of reference data via in memory cache or triggering multiple streams with different processing and stream joins etc.
How did you handle Data Latency requirements?	Multiple options can be considered: 1) move closer to source of data and avoid multiple hops 2) utilize Lambda architecture 3) design and understand eventual consistency 4) predict missing data or design thresholds 5) smaller batches, multiple materialized stores so as
Schema on write can be achieved in Data Lake?	Data Lakes tend more to be schema on read and data warehouse tend more to be schema on write. The schema-on-write data stores require a lot more up-front preparation and ongoing transformation of the incoming data , but depending on your modeling you

Data cleaning and govereance is key to meaningful insight and actions. And often it needs manual intervention to understand and classify. When incoming data is TBs/day, how we can process, clean, store and act in realtime?	Processing TBs/day requires good design of your pipelines / solution and there is no easy way. Based on business needs, you can break the data into smaller subsets and look at required data (for example you might be interested in 20`30 fields out of 1000 odd fields etc.) and can set up some batch jobs to looks correlations / dependencies etc from to other data sets
How the structure data is stored?	You can use any relational databases or even document oriented databases to store and analyse structure data
How retraining of the model is done on the fly based on data being collected so that model can adapt to change in data?	We need to look at concept of supervised learning models and unsupervised learning models / self supervised learning models
What are the opportunities for senior mechanical engineering designers in the filed of AI/ML especially in Industry 4.0? Why are all the industries going towards it?	A mechanical engineering designer has great role in utilizing AI/ML technologies for doing better designs (various simulations under various constraints and rules in efficient ways), new innovations and can combine power of technology with design skills.
Several customers still prefer On Premise for managing data rather than Cloud. Do you have some recommendations for such customers?	Data security might be the prime reason why customers still choose on-premise data storages. We need to explain the importance of hybrid clouds and the trends in utilizing scalable, on-demand infrastructure, pay as you go use in cloud with highly secure set up. Teams should focus on building solutions and not spend a lot of effort in managing
Can you give a case study of "Big Data Analytics" in IoT use case for SMART AGRICULTURE?	There have been lot of solutions build in field of agriculture like automatic soil / moisture testing, detection of pest infestations, monitor crop health, Precision farming (water management, crop rotation
Request you to share OPEN SOURCES available for implementing IoT from Device to Application(End-To-End).	In the session I have covered hadoop based open sources for managing data from sensors to storage like spark/flink / storm / bream, hive, kafka etc.
Pls describe scenario for Data Lake Vs Data Warehousing to co-exist. Is it practical to imagine that a Lake be shrunk to be a warehouse for real time usage	You can think of data lake as storage of all data and data warehouse as having sub-set / processed/transformed data from data lake, so
Flink Vs Spark ? Your view points?	If there is a batch then I would go for Spark, but if it is pure streaming then would go for Flink.
Questions with respect to data warehouse: In Industrial equipment, you have indicated both Data lake and Data warehouse. Does that mean in our big data infra, we need both? What application can we use for warehouse since we are looking at open source solution?	Industrial equipment has both kinds of data, streaming (sensor related) and transactional like work orders, notifications, alerts etc., to manage and gain insights from these both types of data sources you would need a data lake and data warehouse.
What's your platform for DWH?	Since we are from SAP we use SAP DWC and SAP Data Intelligence products
Is it on-prem or On-Cloud?	Depending on use cases, both the options can be implemented.

How does data privacy laws impacts big data like GDPR, etc.?	Data privacy laws have a big impact on aspects of what customer data is collected, how is it processed and where it is stored. So you need to understand what kind of personal data you are handling, sensitivity level, is it required to be anonymised, masked, how do we achieve customer opt-in / opt-out information, what is the retention period etc. factors need to be
Did you use Snowflake or PRESTO - for DWH ?	We used SAP DWC (Data Warehouse Cloud) and some other open sources.
The problem is more pronounced with unstructured data like images, videos, etc., where machine learning may not give us accurate information to act autonomously. How such situations are addressed today for realtime decision making?	Processing accuracy of images and videos has improved significantly, but in general business should build in different processes based on accuracy lets say if it is 90%, then it is automated, but say 50% then you introduce manual processing steps and as the model gets better training content the manual step can be
Is it recommended to use Mongo DB for IoT based applications?	Some customers have used MongoDB as part of the IoT stack.
Please suggest any tools which are Automated and simple to use for Business Analysis. For example, if I have data and I load in the tool, I should be getting Analysis results and recommendations.	There are lot of tools in the market like InfiniteInsight, SAS Enterprise Miner, IBM SPSS Modeler, and Statistica or some open source tools like R package or Weka
How do we ensure data quality?	Quality need to be enhanced in multiple steps --> Data profiling & trust validation of data sources --> Implementing de-duplication stratergies in pipelines --> Auto filling / validations --> data lineage and implement traceability
Any new technology implementation comes with a cost? Are there any open source tools to quickly calculate ROI for faster decision making process?	ROI is not a pure technology driven aspect but more from business, for example how can you measure customer satisfaction & multiplier impact (repeat customer, positive word of mouth etc.) that you were able to achieve due to some prediction.
How is the data getting captured from PLC's from manufacturing shop floor since they are of various types and locked? What would you do with incomplete data since a lot of systems are legacy systems currently?	You can use IoT gateways to connect to PLC's , convert protocols and upload to relevant solutions or utilize OPC connectors.
Is digital twin and virtual twin experience both same?	Digital twins are virtual replicas of physical devices.
What is the industry benchmark on accuracies coming from the ML and deep learning models?	There are some organizations that provide the benchmarks like https://www.eembc.org/mlmark/, https://mlperf.org/press#mlperf-training-v0.7-results
What are those prediction algorithms?	In the session we talked about algorithms to find out Remaining useful life (RUL) and probablity of failure (POF) etc. for machines
In my case, mongo is getting 8 Crore records per day. Does it support that amount of data?	Ideally, it would support such volume, but it will also depends on complexity of document collections and kind of analysis you need to perform on that data.