* **SP24: DATA-228 Sec 12 - Big Data Tech and App**

**Group 3**

**SINDHU NAGESHA - 017419987**

**PRAYAG NIKUL PURANI - 017416737**

**SYED FARAAZ AHMED - 017428619**

**SAI VIVEK CHUNDURI- 017435301**

Question 1a:

Wer have taken interview of one of your senior Harshith Reddy Uppula <https://www.linkedin.com/in/harshithreddyu/> , he is working in tik tok as a data engineer we are following with the interview questions.

1. What about spark and flink will flink take over the market capture in future ?

So like you can't, I would say you can't compare Spark and Flink because Spark is like mostly used in a batch processing, is a batch processing framework. And like a flink is mostly used as a streaming, real-time, framework. So, yeah, I think Spark has, like, Spark has its own advantages and like think has its own advantages and you can't compare both of them. But yes, like Spark has been in has been in the industry. For a very long time and right now and they are like more and more new framework that are coming up, more and more, new, framework now and they're like, more and more, new, frameworks that are coming up, to So like Krano. Presto, at the other latest engines, so that are, that are mostly popular now. I mean, expire is like used in various things like it's used as a query engine and as a framework. Correct so I mean like yeah I think so there are a lot of new technologies I mean like the like I think link is like right now so it's a very popular framework for real-time processing. Okay, it has all the latest functionalities. I'll not dive into details, but yes, so currently like in tech talk so we use both spark and sling. So, Sling, as of Spark is our primary, query engine for all the batch processing. And like the fling, is a real-time processing framework, and, and, and the processing engine as well.

1. What about security and privacy control techniques that are widely followed in the industries ?

Yeah, technology tools and framework for security for big data. Right, so basically so they call it like so in the data engineering term it's called as data governance. But basically you I mean like to you mask the PII information which is personal like PII information to all the employees or like what whoever working so basically yeah like data and governance has been it's very important and critical part so right now I mean like with the boom of AI coming up so that so there has been a lot of analysis is being done with the data. I mean, it's very important to hide, the personal information. Right, so I mean like so, so you as you have been hearing, like you saw, I think you saw the foundational hearing of TikTok, MAY, of TECH TALK is really concerned about this, the user information in the US. So I think, yeah, I think data and governance will place an important role in the US. So I think, yeah, I think data and governance will face an important role in this manner but like when you when it comes to tools and frameworks I don't think there is an important role, in this manner. But like when you when it comes to tools and frameworks, I don't think there is any specific tool and frameworks, I don't think there is any specific tool, for this or like it's, I don't think there is any specific tool for this or like it's just like basically are doing, for this or like it's just like basically, if there is any specific tool for this or like it's just like basically are doing, doing a hashing, doing hashing, of all this personal information and like keeping this hash table separately. basically, right. So, so personally, I, am not, I'm basically, basically, right. So, so personally, I, am not, I'm not in the security and governance team, you know, so that, so that's a completely different team. You know, so that, so that's a completely different team. But like we do, we do, manage like few of the security pipelines but so those so we don't we don't do all the data hashing and masking for the data but like we we serve as a data engineering back-end for the, security and privacy or. Right, so I mean like if so for that risk suppose like if any there are in is there a lot of fake accounts in TikTok right so we manage to detect them and then like, band, band them, like we will pipelines for it. And like, but, but we do not, work on like masking the information or PI information and all the separate team and all but I don't think there is any there is any specific tool of streaming for that but yeah generally generally I'm like, it's kind of a technique to master the information that's it.

1. So tell me about the trends and advancements that you are looking at, in data engineering, field and how are you trying to?

So I think, yeah, I think I've been, so I have. I had worked as in India as a data in use for 1.5 years where, where I, it's parked my interest towards data engineering and I completely, applied data engine roles when I came here. So I think data engineering has been here for a long time specifically big data has been has been here since 2,002 so there has been like a lot of advancements being going on in the time sharing I think, we can discuss this question in 2 parts. So one, I feel as a danger to data engineering and one I feel as an advantage. So the thing I feel, that's affecting data engineering is the kind of new technologies that are coming up tools and technologies. So if you see I mean like also for everything so there is an automated tool coming up. So abstracting upstarting a like underlying technology or framework. So if you see like the Hadoop is like the core component of every distributed are processing right now. I'm like, crossing right now. I think every tool is built on top of them enhancing, its architecture a bit more, right? So I mean like you all have heard about airflow, right? So airflow. So I think airflow has been built to make workflow orchestration or management. So I mean like right now and like there are a lot of tools being like built on a top of airflow that manages airflow and then give you a no port or less code environment like, Daxter, and you know, a perfect oh, something like that. So like what I feel is like if there are like like if people are building these kind of low pod or no code environments. I mean like it's, basically you go to UI and like set up like basically do all the configuration and pipeline will automatically create its own dependencies and run accordingly. So this is like not only about a flow, but like usually for everything. Like when you look at I suppose, any suppose that there are some AI tools coming up. So which can basically if you type in what you need to what you want and it will automatically write a query for you and then I'll give you an, give you an answer. So I'm like these kind of things are like better, like a danger to data engineering I feel because like if you build these kind of environments anyone can do that right there. You I mean like there's no need of a technical expertise in that case. There is a need but like you need to know on other concepts as well other than code.

1. So you have worked a lot on assignments and homeworks back in school ?

So what is that one thing that you have, been utilizing more in your work or that one thing that you're so thankful for. That you learned back at school. Oh, I'm like, what's about the question? So one thing you are utilizing to do my homework or Yeah, let's say, to a tool, maybe even I worked as an, data engineer back in India, but I was not much familiar with link and stuff. I came here and then I'm learning about it here. So something like that that you learned at school and now you are utilizing mode at work. Oh, okay, got it. So, yeah, and definitely, I mean, like, so there was an interesting use case in TikTok. So there, so we have so So basically I'm in that created analytics team. So, I don't know if you guys have used TikTok before. So there so for every video you post you get to see some analytics on it on a data day basis right I mean like basically you have like 1,000 viewers coming up so there will be a particular analytics per view or distribution like so viewers so how many like basically how many mail are watching your video, how many emails are watching your video, like from which kind which, city, these views are coming, coming from, and like for. like something like that. So I, so in last quarter, so I have a feature called, so distributing. So basically distributing this viewer types into view, new viewer and returning viewer and a follower from non So basically you need to, that there are 1,000 US, per if a viewer comes into your profile for the first time. Right. In your profile, he's seeing your profile for the first time, then he'll, he'll be considered some newer new viewers for that video and if he has seen your videos in past any time before he will be considered as a returning viewer in that case. So we were working on this use case. Usually you see, I mean, like there are 1 billion viewers in over TikTok, Daily.1 billion at daily active users and 400 million daily after creators. Okay, so. So, I mean, like if you consider one day of data, so right. And then if you, so usually the point is like the 1, one creator will be averagely having. I think 10,000 views, on, on average. So if you consider that metric, so 4 billion into 10,000. So we want to store a mapping of like. 10 power 8 plus 4 8 10 part 12, turn 10 per 12 rows for each day, right? I'm like it's very difficult. Also, so 1010 per 12 goes for each day.

1. So as a data engineer, you might be dealing with many kind of semi-structured or unstructured data?

So, yeah, I think Yeah, we would be definitely using. Unstructured data, unstructured and semi-structured data as a data and there but right now so we do so currently i personally we like most most of the 80% of the work is being done using structured and semi-structured data. Only 20% of the work is currently done using unstructured data. I'm like that's been handled more right now. It's been handled more by a machine learning engineer rather than a data engineer. Right. Okay. I mean, like, that's what I said in the beginning. In the future, I'm like, data engineering will, will be handy, will come in this picture a lot to do to deal with this unstructured data. So when, dealing with this, Mindy, dealing with this semi structure data basically. So this semi structure data will be so if you are so you have working over a bigger organization usually your data will be stored as documents, document documents like JSON, JSON files or JSON files mainly, right? So, so it will be no, no, SQL database, correct? So, and then basically we what we'll do, we will use that semi-star on semi-structured data and then we pull into a hype or some data warehousing architecture and then we convert into a structure format.

22:29:57 Right, so basically if you for example, so if you go on to like tick to over any app, so you do what a source of was you watch a video or you like a video or you come into video for everything every every user activity is It's collected in 2 locks. So those logs will be a very big. Like Jason for like document file for format. So you basically convert those documents into meaningful information like into structured format when you bring to hive you do some processing and convert into a structured format and then ugly apply the sequel for each to get meaningful information out of it. So we personally do not, I mean like as a back-end data basis, there are like lot of databases like a base, a base h base, Mongodb. These kind of databases we use as like no sequel databases. So these are all like online databases. Or like the backend databases and then we basically extract those the extract that data into HIV and then do some processing, get some meaningful information out of it and then send back to these databases.To show them on the on the app like the analytics. Both analytics or like the likes and comments you see the number of likes and comments and number of sheets that you see to those kind of things happen.

1. So next question is what is the most challenging, project or task for you?

So we are trying so we are trying to track the interest of your followers. And we are trying to show you some of the top creators from you from your followed preferences. So that you can inspire from them or you can learn from their videos. So we are working on those recommendations system we are trying to recommend 1010 creators and 10 items on a daily basis to the creators. So this basically will these craters and items will be on will be in same category that you're posting videos. Suppose you are a travel, travel video creator or suppose you are a cooking video creator. So we recommend these travel. Videos and travel creators and cooking creators to you. Right, so, so imagine so suppose it's a it's a graph problem. So basically if you are a creator you will be having 1,000 viewers. Suppose you have like suppose you have 2 viewers. Next up both and those viewers in each year will be seeing 10 videos. 10 videos. And then, so basically they are seeing 10 other creators each each your viewer will be creating 10 other creators for the same. The second year also will be seeing 10 other creators, right? It will be like 20 creators for a person, for a creator with only single video and only 2 viewers. To imagine there are 400 million creators and then each so every day there is like 600 million videos being posted and 1 billion daily active viewers. And those 1 billion daily active viewers will be viewing other creators, other 400 million creators. And then it's a very, it's a very complex problem, you know. I'm like, so I think it's, so I think it's, so I think it's, very difficult to solve.

22:34:23 And, so we usually use HIV as I said. So we usually use HIV as I said. So we usually use HIV as I said. So we usually write equal So for this, so we need to come up with, some kind of, so that, basically, a recommendation algorithm, right? So we do not use an ML, but it's kind of an algorithm. We, developed a UDF, which is usually defined function. So which, basically, so that, that's a new whole set of algorithm and then, and that basically makes your work easy and then, it, Yeah, we use all the complex data types in in SQL like, maps and arrays and then make sure the data storage will is list and accomplish the complexity is also reduced. Right now I think I'm currently working on that. I think it will be deployed. This quarter. So I think using this, so keep coming, so using this like complex architecture, I think we have a greatly reduce the complexity of the problem.

1. So as you have worked as a data engineer in both India and US ?

I think, so you usually what I hear is the Usually what I hear, so basically, so I can't compare my work experience because I work for a service based company in India and I'm here I'm working for a product based company.

22:35:54 And so unlike that it does the scale of data is completely different and the use of tech and tech is completely different. So I personally cannot give, give a justifying answer for this, but usually I'll tell you what I hear, but in usually what happens is suppose you have a product. So, you, so suppose, so you're working on, so you're working for a company called Uber. Okay, so or Amazon. So most of the core teams, most of the core kind of work is being done in the US and not in the India and not in India. And it's not always, it's not always correct. Mostly it is like that because, you need to, you need to be in the same container same time. Zone to answer critical questions and like do like for collaboration and not with collaboration with the with senior, management and also you need to be in the same, region. So that's more important. So that's why more, More core teams are in US. I would say, but it's not true. I mean, like so. So I know you think like basically if if we like if I'm working on TikTok I'll so I'm like not so let's let's let's like if a person is working in open AI or if I person is working in meta, you you think like he's earning really really a lot like really, very much more than your ex, but like it's not really it's not not always true. It depends on the team he's working in. I mean like if it's a core team and it's an important team then he will be, he, he will be paid more. There will be like more, there will be more important teams in India, some important teams in India, some important teams in India, which has, more, which makes more impact than some teams in US. Right, so. I think that's what that's the difference. I mean like the tools and the tools and technologies that you that you use basically differ from team to team. I think the second, the second perspective for this question is like, suppose the latest. That you suppose, it, a unique tools comes up, like, if suppose a person has, has built some technology or, support automated tools. Suppose like, a suppose, let's support Daxter. So it's an it's an automated service of airflow.

22:38:11 So basically mostly people come to these conferences in Bay Area to basically to promote their tools. So here they meet with all the people like suppose. They met with the person working on tic talk so they they what they do is they usually promote their tool. So they, what they do is they usually promote their tool, they usually promote and then they try to, they try to use this and they try to promote their tool such that these people use it in that way i mean like the latest technologies and technologies will be basically mostly firstly, inhibited by the US team for the US companies and then it will move on to India. But it completely depends on various according to the company or you are working and the team you are in the road, role you are in.

1. We hear data engineering term, there is something called data structures which everyone specifies it ?

So I just wanted to know how important is data structures in data engineering. Yeah, I mean, to be honest and I'm very bad at data success and like I hate data just so that's the reason that's the reason I want to like be in data engineering. In the first place during also when I was in India. So yeah I think so like to 2 parts like one is for the end. So one is for interviews and one is, one is like real work. When it comes to interviews, definitely yes, you need data structures and you need to practice legal code. You need to do those those important hit code questions like the first 100 questions or 300 you called questions that you need to do when you don't have some basics or understand. I will I do. I don't say that you need to be so, you need to Give the optimized version of the code. If you could, write a code that, that will, solve some test cases. That would be enough. Right so as a as a data engineer that would be enough. Secondly, when it comes to work, I think, 90% of the time you don't use this data success. Okay. But I should think I have to think the same. I used to, I'm like, I'm, I, in my previous years of experience, I never use data structures, right? But like in a while when I came to TikTok, but I used a lot of data structures and like lot of computer science things, in my, you know, day-to-day day-to-day work here, you know, to deal with this big data, big data and big volume of data and not to optimize them. Yeah, I think yeah it definitely helps. It definitely helps but if you are starting right now. But I would not suggest you to spend much time on it. So if you have some skills from before it definitely adds up to your profile. But yeah. Thank you. So my next question is, as you know, we have a big data subject and, we have to create projects. So I just wanted to know which tools you would recommend or the current booming industry we can implement so that it can be helpful even for the industry and also for our knowledge. Yeah, so I think so I would not comment on like the project that you can do, but I would, I would give you some tools and technologies that you can, learn. Right now, so that it will be easy in your interviews and then when you come to industry to work right so 1 one like basically Hadoop is like is the building blocks for every architecture that we have now. My first suggestion is like try to learn this underlying technologies, right? So suppose if you are learning about. Redshift if you learn learning about BigQuery. Like you, you usually mentioned like I have used that shift I have used BigQuery but like how does it work?

1. How what's the underlying architecture of it, architecture of it is the important thing that you need to know, right?

The concept behind it like how they built it and what the what's the purpose of building it and what the walker what kind of architecture are using are they using behind just not like learn how to so you learning like how to use the UI off redshift is very easy but like learning how they implemented it and what's underlying underlying extracted, code or technology behind it is very difficult, but like that you need to

1. So you said Azure and then AWS as well so, it ok if we focus on only one?

So, they are 2 different things. So I heard from people saying that if you learn one cloud technology that should be suffice to, you know, learn the other one. So, is that really helpful or we should have a basic idea of all the in cloud platforms? Yeah, definitely. I'm like if you have a core idea about like one of the cloud technology, the cloud services or cloud technologies. Yeah, it's about like one of the cloud, the cloud services are cloud technologies. Yeah, it's efficient because only mostly, the competitor, the competiting tools like big, and redshift, like, the competing tools are like BITQUE. And, as you have synapse and analytics, right? So those are like those will differ in like some, those will different in architecture in some way. I'm like some can there'll be some kind of difference but the core area is seen the core idea will be very safe. So I think if you you can you can focus on like one cloud and then basically learn in and out of it and then rest all will be it'll and after that it's it's very easy to learn the other things it's just like you need to know the difference between the other things. It's just like you need to know the difference between these 2, these 2 things. But the code is very similar. Yeah. I think that's sufficient. Yeah, I think. I think that's true. I think you can deep dive into one cloud and then and arrest like the competitor, the competitive services in other cloud services are pretty much similar.

Question 2b:

<https://drive.google.com/file/d/1Xe_4B1v818u341aC1JUh2-FF062QiX3S/view?usp=sharing>