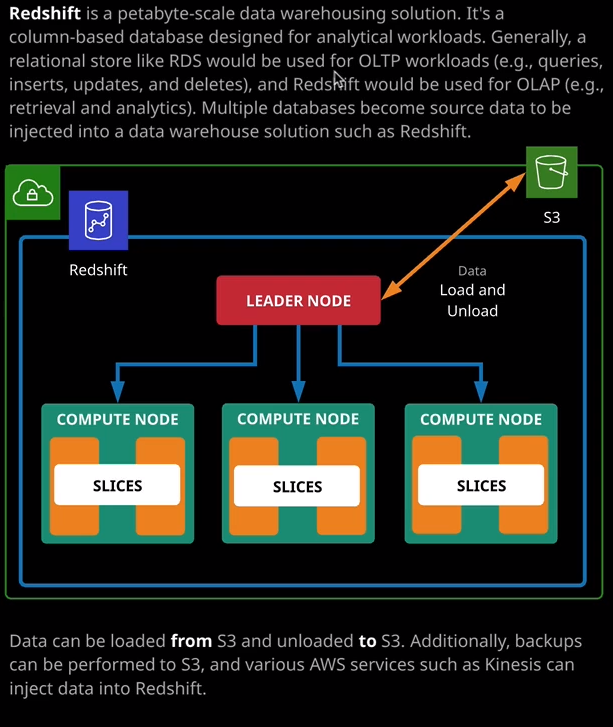
**Red Shift is a petabytes scale data warehousing solution from AWS. It's a database designed for OLAP based transactions, which stands for online analytical processing and this is compared with OLTP which is online transaction processing, which is what most normal databases do**. Now Red Shift isn't the type of database where you would add and update individual records. For that, you would use something like RDS. **Red Shift if used for data warehousing and analytics. It's a column based database, and that means that data is stored in columns, so everyone's first name is store together. Everyone's surname is stored together, everyone's age, and everyone's address. They're all stored together in the same location on physical storage devices. So column based databases are better at doing queries which look at large amounts of the same type of data**. So if I wanted to query all students inside the Linux Academy database doing the query which look for any patterns in ages, or locations or looking for everyone with a certain surname or counting the number of people with a certain surname then a column based database would be much more efficient than a row based database.   
A row based database is where every single row is a unique record. So you would have a record in the Linux Academy database, and I would have a record and if you were performing queries or operations, which interacted with individual records. So updating your address or updating your age or changing the courses that you're currently studying then a row based database would be much more efficient. So we generally use **row based databases for transaction style processing and column base databases for analytical processing**. Now traditional data warehouses tend to be expensive to provision and operate requiring time, infrastructure, and skills.

**Red Shift, though, could be provisioned on ad hoc basis for a particular task or only used when you require that warehousing functionality. So Red Shift uses a cluster architecture, and it's capable of loading and unloading data to and from S3. Red Shift can be used as a target for many other products across AWS as a final data storage location, for example, Kinesis and Kinesis Firehose** that I've talked about in a previous lesson.



Architecturally, a Red Shift cluster has a leader node and a varying number of compute nodes, which actually perform the distributed querying. Each compute node has slices and slices are what queries are actually computed on, so a compute node may have a very number of slices and when the leader node decides to execute a query, it splits it up into individual components and allocates that component of the query to a particular slice running on a compute node. When you connect to Red Shift using any querying tools, it's the leader node which handles these connections. When you're answering questions in the Solutions Architect Associate exam, you do need to be aware of what Red Shift actually is. You need to know that it's a data warehouse. You need to know that it will generally function as the target data store after transactions have been processed in other stores. So data that previously existed in the Kinesis Firehose might make its way to Red Shift. Data that's been used inside RDS for real time transactions that might make it into Red Shift. Red Shift is used for analytical or summarization transactions and that's because it's a column based data store, so it can scale to almost any load levels that you need and it's designed to operate at the petabyte level. So it's designed for really huge scale data warehousing scenarios. So that's what you need to remember for the exam. Really, it's the when and where to use Red Shift that's important.

Now again, comparing Red Shift to EMR and Athena and knowing the difference is really important. So **Athena is used for doing transactional type queries on data that's in S3. So you'd use it to query the same type of data that you might query with a relational database system but with Athena, you don't need to maintain database infrastructure, and you don't need to load the data into the product first. It can query it directly from S3.   
EMR is going to be used when you need to perform analytics and actual modification on that data. So if you're performing a large scale analysis of semi structured or unstructured data you need to modify it, change it from its original form into something new, that's a use case when you might use EMR  
 but if you're wanting to have an end state repository your data when data's been processed when it's been run through transactional data bases, when you want one single location to be able to do summarized or analytical queries on all of your data then Red Shift is probably that product**. A large organization such as amazon.com, for example, might have hundreds or thousands of isolated databases around the organization there might be databases for customers, there might be databases for orders from amazon.com, for audio book purchases, for videos, delivery information, or purchases from Amazon Prime. There's a lot of individual use cases for isolated databases, but what you might want to do with an organization of that size is to take all of those individual databases, which get used for transactions on a constant basis and then take that data at a point in time and funnel it into Red Shift. You've got all you dated then in one place that you can use to perform some analytical style queries generally for things like reporting, so end of month processing, long term strategic trending, that kind of thing. That's what you're going to use Red Shift for. So generally, the data that's in Red Shift is also going to be in other databases. You're not really duplicating it because the functionality of Red Shift is different. It's designed for a different use case than relational databases or transactional data bases. So you'll have the original data in all of you transactional stores. That's where you'll interact with it on a daily basis but then it would be brought into Red Shift for your periodic processing analysis and trending.