What are linux distributions that supports AI:

- **Ubuntu**: A popular and user-friendly distro that supports many AI frameworks and tools, such as TensorFlow, PyTorch, Keras, etc. It also offers a cloud service called Ubuntu AI that can help you deploy your models easily.
- **Fedora**: A community-driven distro that is known for its cutting-edge features and security. It has a special edition called Fedora Scientific that includes many scientific and mathematical software, such as R, Octave, Scilab, etc.
- **Arch Linux**: A minimalist and customizable distro that follows the principle of “keep it simple”. It has a large and active user community that maintains many packages for AI, such as Anaconda, Jupyter, Theano, etc.

There are other Linux distros that can also be used for AI, such as Debian, CentOS, OpenSUSE, ..etc

If you are using Windows 10, you can also use the **Windows Subsystem for Linux (WSL)** to run Linux applications on your Windows machine. This can give you access to the Linux tools and libraries for AI without having to dual-boot or use a virtual machine.

Top AI AWS :

- **Amazon Comprehend**: A natural language processing (NLP) service that uses deep learning to analyze text and extract insights, such as sentiment, entities, topics, key phrases, etc. It also supports custom classification and entity recognition with your own data and labels.
- **Amazon Rekognition**: A computer vision service that uses deep learning to analyze images and videos and detect objects, faces, emotions, text, scenes, activities, etc. It also supports custom labels and face recognition with your own data.

- **Amazon Lex**: A conversational AI service that helps you build chatbots and voice assistants that can interact with your customers and users using natural language. It integrates with Amazon Polly and Amazon Transcribe to provide speech recognition and synthesis capabilities.
- **Amazon Personalize**: A recommendation engine service that helps you create personalized experiences for your customers and users based on their behavior and preferences. It uses deep learning to learn from your data and generate real-time recommendations for products, content, offers, etc.

Popular Cloud Service Providers:

- **IBM Cloud**: A hybrid cloud platform that combines public cloud, private cloud, and on-premises infrastructure. It offers services for AI, blockchain, data analytics, IoT, security, etc.
- **Alibaba Cloud**: A leading cloud provider in China and Asia that offers services for e-commerce, gaming, finance, media, healthcare, etc. It supports multiple languages and regions and has a global network of data centers.
- **Oracle Cloud**: A cloud platform that specializes in database and enterprise applications. It offers services for data management, business intelligence, integration, security, etc. It also supports Oracle's own products, such as Oracle Database, Oracle Fusion, Oracle E-Business Suite, etc.
- **Salesforce**: A cloud-based software company that provides customer relationship management (CRM) solutions and other business applications. It offers services for sales, marketing, service, commerce, analytics, etc. It also has a platform called Salesforce Platform that allows developers to build custom apps.

13 Vs of Big Data (They became 17):

Source: IRJET-V4I957.pdf

- **Volume:** The amount of data generated is increasing at an unprecedented rate, and this is one of the defining characteristics of big data.
- **Velocity**: The speed at which data is generated is another important characteristic of big data. With the rise of the Internet of Things (IoT), data is being generated faster than ever before.

- **Variety:** Big data comes in many different forms, including structured, semi-structured, and unstructured data. This variety makes it difficult to manage and analyze big data.
- **Veracity:** Veracity refers to the quality of the data. Big data can be messy and incomplete, which can make it difficult to draw meaningful insights from it.
- **Value**: The value of big data lies in its ability to provide insights that can help organizations make better decisions. However, extracting value from big data requires sophisticated tools and techniques .
- **Validity:** Validity refers to the accuracy of the data. Inaccurate or incomplete data can lead to incorrect conclusions and poor decision-making.
- **Volatility:** Volatility refers to how long the data is relevant for. Some types of data lose their value quickly, while others remain relevant for a long time.
- **Visualization**: Refers to the process of creating visual representations of data to help people understand it better.
- **Vagueness**: Concern the reality in information that suggested little or no thought about what each might convey.
- **Virality:** It is defined as the rate at which the data is broadcast /spread by a user and received by different users for their use.
- **Venue:** Various types of data arrived from different sources via different platforms like personnel system and private & public cloud.
- **Vocabulary:** Data terminology likes data models and data structures.
- **Viscosity:** It is a time difference the event occurred and the event being described

What is Internet port:

A port is a virtual point where network connections start and end. Ports are software-based and managed by a computer's operating system. Each port is associated with a specific process or service.

Ports allow computers to easily differentiate between different kinds of traffic: emails go to a different port than webpages, for instance, even though both reach a computer over the same Internet connection .

Ports are standardized across all network-connected devices, with each port assigned a number. Most ports are reserved for certain protocols — for example, all Hypertext Transfer Protocol (HTTP) messages go to port 80. While IP addresses enable messages to go to and from specific devices, port numbers allow targeting of specific services or applications within those devices

What are the different port numbers?

There are 65,535 possible port numbers, although not all are in common use. Some of the most commonly used ports, along with their associated networking protocol, are:

Ports 20 and 21: File Transfer Protocol (FTP). FTP is for transferring files between a client and a server.

Port 22: Secure Shell (SSH). SSH is one of many tunneling protocols that create secure network connections.

Port 25: Historically, Simple Mail Transfer Protocol (SMTP). SMTP is used for email.

Port 53: Domain Name System (DNS). DNS is an essential process for the modern Internet; it matches human-readable domain names to machine-readable IP addresses, enabling users to load websites and applications without memorizing a long list of IP addresses.

Port 80: Hypertext Transfer Protocol (HTTP). HTTP is the protocol that makes the World Wide Web possible.

Port 123: Network Time Protocol (NTP). NTP allows computer clocks to sync with each other, a process that is essential for encryption.

Port 179: Border Gateway Protocol (BGP). BGP is essential for establishing efficient routes between the large networks that make up the Internet (these large networks are called autonomous systems). Autonomous systems use BGP to broadcast which IP addresses they control.

Port 443: HTTP Secure (HTTPS). HTTPS is the secure and encrypted version of HTTP. All HTTPS web traffic goes to port 443. Network services that use HTTPS for encryption, such as DNS over HTTPS, also connect at this port.

Port 500: Internet Security Association and Key Management Protocol (ISAKMP), which is part of the process of setting up secure IPsec connections.

Port 587: Modern, secure SMTP that uses encryption.

Port 3389: Remote Desktop Protocol (RDP). RDP enables users to remotely connect to their desktop computers from another device.

Data Engineering Tools:

source:[10 Essential Data Engineering Tools and How To Use Them (springboard.com)]

1. Apache Kafka

Apache Kafka is mainly used for processing and building data pipelines in real-time.

It's mostly utilized in industries with a heavy and constant data flow that involves analyzing website activity, collecting metrics, and monitoring log files.

Kafka's ability to handle massive volumes of data stream non-stop is the reason a lot of app and website developers use it. The platform will most likely remain in use for years to come. While Apache Kafka isn't easy to learn, it's used by more than 30% of Fortune 500 companies, making it a great time and money investment for data engineers.

2. Apache Airflow

Apache Airflow is an open-source data engineering tool. The main advantage is its ability to manage complex workflows. Being open-source, Airflow is completely free to use and constantly receives community upgrades. With more than 8,000 companies using Airflow to some degree in their operations—like Airbnb, Slack, and Robinhood—it isn't likely to be replaced.

Luckily, it's extremely easy to use. To showcase your skills and abilities, you can build a smart ML model to transfer data and manage a fluctuating workflow.

3. Cloudera Data

Cloudera is a cloud-based platform for data science, machine learning and data analytics. Cloudera Data in particular is popular among large-scale companies thanks to its dual nature, allowing data engineering and analytics teams to use the platform through the cloud and on-premise.

Cloudera has a user-friendly interface and a plethora of tutorials and documentation. It's mostly used by financial institutions like the Bank of America and the Federal Reserve Bank.

4. Apache Hadoop

Hadoop, instead of being a single tool with a limited number of features, is a collection of open-source tools made to manage large-scale data often produced by large computer networks. What makes it a household name for many corporations is its ability to store data in an orderly manner, perform real-time data processing, and provide detailed and clean analytics.

While Hadoop's dependence on SQL for its databases makes it easy for anyone with a background in SQL to break in, mastering the tool would require a lot of time and effort. Hadoop isn't going anywhere soon, especially with companies like Netflix and Uber—alongside 60,000 others—showcasing why it's an invaluable tool.

5. Apache Spark

Apache Spark is another open-source data engineering and analytics tool. While it doesn't have a wide variety of features and capabilities, it's one of the fastest data managing and stream processing frameworks. Spark can queue more than 100 tasks in-memory, leaving data scientists and engineers free to accomplish more critical tasks. It's also compatible with numerous programming languages such as Python, Java, and Scala.

As long as you're keeping your work simple, Apache Spark is easy to use and offers high-performance data processing in a variety of industries ranging from retail and finance to healthcare and media.

However, for more complicated tasks, Spark can add an unnecessary layer of complexity and difficulty. Spark's work model is still finding its way into a lot of useful ecosystems, such as Hadoop, and doesn't seem to be going away anytime soon.

SQL VS NOSQL: