Question1

Answer1:

Conceptual pipeline

Using open source tools:

We could use Hadoop HDFS /Hive storage for colleting data and **Apache spark** to do ETL transformations & analysis and store the tables/views in Hadoop. We can visualise the user website analytics using **Matom**a(which is easier to use and has in built functions useful for bookly analysis tasks). **Open web analytics** is another tool which can be used but it has to be hosted using JavaScript or PHP.

Using Google cloud tools(GCP):

Data can be stored in google cloud storage and can be processed, analysed and visualised using Google analytics. for advance dashboards you can use Looker. Google analytics with python can generate customised metrics which can be used in analysis.

Using Azure tools:

Data can be stored in data lake, ETL and analysis can be done by Pyspark in azure databricks and can be visualised using Azure Application Insights(easier to setup) or PowerBI (with some customisations for web analytics).

Using AWS tools:

AWS Glue for data collection and ETL, aws kinesis for web analytics (easy to use), we could also use amazon athena with sql queries( for querying data and analysis). Amazon Quicksight for visualisations.

My choice would be to use open source tools if we are budget constraint else bookly could use GCP tools mentions which is easier to setup and takes less time to implement, do the analysis and create required visualizations.

Additional Details:

To perform the above analytics using the data pipeline described above:

1. List all-time top rated books and trending ones: The data collected from the marketplace for selling books could be used to calculate the average rating for each book. The average ratings could be calculated on a daily, weekly, or monthly basis and stored in the data warehouse or data lake. The data could then be queried using SQL-like queries to identify the all-time top rated books and the books that are currently trending.
2. Measure user sign-up rate over certain periods (weekly, quarterly, etc): The data collected from the user sign-up process could be used to calculate the number of new user sign-ups for a given period (weekly, monthly, etc.). The sign-up data could be stored in the data warehouse or data lake, and the number of sign-ups could be queried using SQL-like queries to generate the desired sign-up rate.
3. Show the total number of real-time (current) page views for any given book description page (product page): The data collected from the product pages could be used to calculate the number of page views for a given book description page. The page view data could be stored in the data warehouse or data lake, and the number of page views could be queried using SQL-like queries to generate the desired real-time page view count.

These analytics results could be visualized using data visualization tools, such as Tableau, Power BI