PySpark - Practice Sheet

10 conceptual Q&A Data Engineering

Q1. What is PySpark and what problem in Data Engineering does it solve?

A: PySpark is a tool or platform that automates or simplifies a specific concern in the Data Engineering domain, reducing manual effort and operational complexity for engineering teams.

Q2. What are the core components or concepts in PySpark?

A: PySpark has key building blocks that work together: a configuration layer defines intent, a runtime layer enforces it, and an API or CLI layer allows operators to manage and observe the system.

Q3. How does PySpark compare to its main alternatives?

A: PySpark trades off simplicity against feature richness differently than its alternatives. Choose PySpark when its specific strengths performance, ecosystem, or ease of use align with your team's primary constraints.

Q4. How do you get started with PySpark for a new project?

A: Install PySpark via its package manager or binary release. Follow the quickstart to initialize a project, configure the minimal required settings, and run the default command to verify the setup works end-to-end.

Q5. What configuration options most affect PySpark's behavior in production?

A: Production-critical settings typically include concurrency limits, timeout values, retry policies, resource limits, and logging levels. Review the official production hardening guide before deploying.

Q6. How does PySpark handle reliability and high availability?

A: PySpark provides HA through leader election, replication, or stateless horizontal scaling. Deploy at least two instances across availability zones and configure health checks for automatic failover.

Q7. What observability does PySpark provide out of the box?

A: PySpark exposes metrics (Prometheus endpoint or built-in dashboard), structured logs, and optionally distributed traces. Define SLOs on the key metrics and alert before users are affected.

Q8. What are common performance bottlenecks in PySpark?

A: Performance bottlenecks typically stem from misconfigured connection pools, large payload sizes, insufficient worker threads, or missing caches. Profile with built-in diagnostics before optimizing.

Q9. What security best practices apply when running PySpark?

A: Run PySpark with the principle of least privilege, enable TLS for all communication, use a secrets manager for credentials, and keep the software patched to the latest stable release.

Q10. What mistakes do teams commonly make when adopting PySpark?

A: Common pitfalls include skipping the documentation and learning the API by trial and error, using default configurations in production, lacking a runbook for operational issues, and not testing failover scenarios.