**Data Pipeline & Analytics Architecture Blueprint**

**Title:** *Scalable Data Ingestion, Storage, and Analytics Architecture for Smart 5G*

**A. Overview Diagram**

(Visual architecture diagram showing flow from data sources to analytics dashboards)

* Data Sources → Ingestion Layer → Stream Processing → Storage → Analytics & ML → Visualization & Alerts

**B. Architecture Components**

| **Layer** | **Function** | **Technologies** |
| --- | --- | --- |
| **Data Sources** | Network elements (RAN, Core), OSS/BSS, Telemetry probes | Kafka Connect, SNMP, gNMI, NetFlow collectors |
| **Ingestion Layer** | Real-time and batch ingestion | Apache Kafka, AWS Kinesis, Fluentd |
| **Stream Processing** | Filtering, enrichment, anomaly detection | Apache Flink, Spark Streaming, Apache NiFi |
| **Storage Layer** | Scalable storage for time-series and unstructured data | Cassandra, HDFS, AWS S3, InfluxDB |
| **Analytics & ML** | KPI computation, predictive analytics, model training | Spark MLlib, TensorFlow, PyTorch, Kubeflow |
| **Visualization** | Dashboards, alerting, reporting | Grafana, Kibana, Tableau |
| **Orchestration** | Workflow and pipeline orchestration | Apache Airflow, Argo Workflows |

**C. Key Design Considerations**

* **Scalability:** Elastic scaling to handle bursts in network telemetry and user data.
* **Low Latency:** Near real-time processing for critical KPIs and anomaly alerts.
* **Data Quality:** Ingestion pipelines include validation, deduplication, and schema enforcement.
* **Security:** Encrypted data in transit and at rest, access control at every layer.
* **Fault Tolerance:** Replication and failover mechanisms to prevent data loss.
* **Extensibility:** Modular architecture to add new data sources and analytics models without downtime.

**D. Example Data Flow**

* **Raw Data Capture:**  
  Network probes emit telemetry → Kafka topics for ingestion → Spark Streaming filters and tags data.
* **Processing & Enrichment:**  
  Correlate events from multiple sources, timestamp synchronization, anomaly detection.
* **Storage:**  
  Time-series KPIs stored in InfluxDB; historical logs archived in HDFS/S3.
* **Analytics:**  
  Batch ML training on historical data; real-time prediction of congestion and faults.
* **Dashboard & Alerts:**  
  Grafana dashboards visualize throughput, latency, and energy usage; Prometheus triggers alerts for SLA breaches.