**Data Sources & Ingestion Profile**

* **Sensor projects**:
  + *Machine Utilisation*: 1 Hz data, 10–15 projects, typically 30–45 sensors.
  + *Air Quality*: 1/30 Hz data, ~10 projects, typically 10–15 sensors.
  + Total cap: ~60 sensors across all projects.
  + Data runs continuously (24×7, including weekends).
  + At full scale (Month 6), daily volume will be ~2.6M–3.9M rows (~105–157 MB compressed per day, ~3–5 GB/month).
* **Intensive projects**: 10–15 companies, each providing structured (CSV/XLSX/JSON) and unstructured (PDF/PNG) files.
  + Batch cadence: weekly to monthly, with occasional historical backfills.
  + Per-company volume: 5–50 GB, giving a total ~50–500 GB across the cohort.

**Rollout Timeline**

* The programme runs for 6 months and will ramp up gradually.
* *Month 1*: platform setup, test data only.
* *Month 2*: 2–3 pilot SMEs onboarded.
* *Months 3–4*: 5–10 additional SMEs.
* *Months 5–6*: ramp-up to ~30–40 SMEs total, finishing with 10–20 active projects.
* This staged approach means compute/storage needs will scale progressively, not all at once.

**Data Transformation**

* Mainly structured/semi-structured (CSV, JSON), with some unstructured files stored for reference.
* Transformations: schema validation, deduplication, enrichment, rollups (minute/hour/day), and modelling into tables for BI.
* Frequency: near-real-time (1–5 min) for sensors, batch-on-arrival or daily/weekly for files.

**Outputs**

* BI dashboards: ~60–120 total across SMEs
* Users: ~20–40 named users, typically 2–4 per company, with ~10 concurrent peak.
* Latency expectation: <5 minutes for key sensor KPIs, seconds for dashboard queries.

**Data Science / ML Workloads**

* Light ML and analytics: anomaly detection, forecasting, utilisation metrics.
* Workloads: feature refresh daily/hourly, model training weekly, scoring daily.
* Scale: small to moderate; exploration + batch scoring.

**File Types**

* Structured: CSV, XLSX.
* Semi-structured: JSON.
* Unstructured: PDF, PNG (metadata managed in Snowflake).

**Other Notes**

* We’ll use role-based access per company/project.
* Warehouses: small/medium with auto-suspend/resume; separate read warehouses for BI.
* Raw sensor data is retained for ~6 months; rollups are retained for ~12 months.