**What the Dataproc Sub-Agent does (at a glance)**

1. **Collects signals**
   * **Cloud Monitoring** (cluster/YARN/HDFS/job metrics) [Google Cloud+1](https://cloud.google.com/dataproc/docs/guides/dataproc-metrics?utm_source=chatgpt.com)
   * **Cloud Logging** (driver + YARN container logs) [Google Cloud+1](https://cloud.google.com/dataproc/docs/guides/logging?utm_source=chatgpt.com)
   * **Job metadata** via **Dataproc Jobs API** (state, timings, UUIDs, appIds) [Google Cloud+2Google Cloud+2](https://cloud.google.com/dataproc/docs/reference/rest/v1/projects.regions.jobs?utm_source=chatgpt.com)
   * **Spark event logs** (JSON) from **Spark History Server / event log dir** for stage/task metrics [Apache Spark+2Apache Spark+2](https://spark.apache.org/docs/latest/monitoring.html?utm_source=chatgpt.com)
2. **Builds a performance memory**
   * Daily facts in **BigQuery** (job runs, stage stats, resource usage).
   * Baselines (p50/p95) per job/app over trailing N days.
3. **Detects anomalies & pinpoints causes**
   * Idle or unhealthy clusters; over/under-utilization.
   * Jobs exceeding baseline (e.g., *“avg 10 min; today at 28 min”*).
   * Stage(s) driving delay (e.g., *“Stage 4 takes 18/28 min, heavy shuffle & spill”*).
4. **Acts (guard-railed)**
   * Alert owners with evidence + recommended fixes.
   * Open tickets; (optional) autoscale/stop non-prod clusters with **graceful decommission**.