

Entity Resolution Service – Service Definition Document, version 2.0

Overview of the Entity Resolution (ER) Service

The ER service enables the resolution of entities (organisations or people) from a variety of records, where identifiers found in these records are used to point to the existence of an entity in the real world. These identifiers are then stored against the entity to enable users to understand how an entity's identity in data is composed. The service uses a variety of largely statistical, and so explainable, techniques to assert the existence of an entity, but can also indicate where there is insufficient data to do so in order to minimise false positives.

Records used in the entity resolution service must contain certain identifiers (e.g. names, addresses, IDs, phone numbers, etc.) for the service to work. The service is particularly adept at resolving human entities as it is optimised to work on the messiest identifiers – names and addresses. Furthermore, the system has been extended to work with various types of events, including travel and commercial/financial data (e.g. purchase orders).

The service can work at billion row scale (e.g. bulk data). It is a real time system.

The service requires a variety of software components to run in the Cloud, including: Sensing Entity Resolution engine API; Elastic Search index.

A variety of data visualisation tools are provided which extend the entity resolution service. These include: network charts (graphing); calendars; and geo viewers. These tools enable analyst users to see connections, spot patterns, and visualise events. Advanced querying is available. An alerting module is also provided; this provides notifications when changes about a specific entity's identifiers (e.g. a new phone number) are detected in large scale datasets.

Deployment model: Cloud Software as a Service (SaaS)

We provide a mixture of proprietary, open source and Crown-IP applications running on a cloud infrastructure run by either the client or by their cloud service provider. We will manage (install; set up; configure; run; maintain) all entity resolution related services and tools.

Information assurance "C Impact Level (IL) at which the G-Cloud Service is accredited to hold and process information

As our services are run in the Client's cloud, the accreditation level is dependent on Client's cloud and the requirements of the datasets to be exploited by the service.

Details of the level of backup/restore and disaster recovery that will be provided on-boarding and off-boarding processes/scope etc.

As our services are run in the Client's cloud, accreditation, back up, DR etc. is dependent on the Client's cloud and associated services available therein.

Service constraints

Maintenance windows are dependent on software upgrades. These will be ad hoc and are not scheduled. Customisation of the stack (e.g. feature extensions to the tools) can be provided but this will be at additional cost and via mutual agreement. The service requires specific hardware; the specification of the hardware depends on the data volumes to be processed.

The service requires certain datasets to be made available. It is the Buyer's responsibility to ensure that such data is available (we can advise on the selection and suitability of datasets once the requirements are fully understood.)

Service levels

Operational service availability is dependent on client needs given that data engineering support will generally be required to run the service.

We are able to provide support services within normal working hours, i.e. 0900-1700, Monday – Friday. Due to the highly technical nature of the analytical services, we only offer a blended technical support package that includes all first, second and third line support. Once the buyer's specific support requirements are understood, we will be able to agree recompense arrangements should support levels not be met.

Associated data engineering and analysis services

Dependent on the buyer's aims and objectives, we can provide a variety of advanced data engineering services. These cover dataset evaluation, preparation, transformation, ingest and load related tasks.

Data analysis services are also available. These cover: advice on dataset selection; insight generation; and investigation of entities of interest. (See separate SFIA rate card for rates.)

Training

We offer online introductory training modules that cover the basics of entity resolution, the use of the visualisation and query tools as well as how to deploy a simple ER service. We can also offer bespoke in-house training; this training can be tailored to the level of skill and expertise possessed by the Client and their specific needs.

Business change, consultancy services

Dependent on the delivery model, we can offer business change services. These aim to ensure that new ways of working consistent with the overall service offer are embedded to maximise the chances of successful delivery where the client elects not to outsource all services to us. These services embrace process development, policy formulation, change planning and communications. (See pricing document).

Ordering and invoicing process

Due to the bespoke nature of the service, the Buyer will need to advise us of their specific needs via a formal service request. These will include: client's aims; the datasets they have available; hosting deployment options; business uses cases and the service outcomes required.

Set up, configuration

We can set up, configure and install all required services in the buyer's cloud environment. (A costed technical evaluation will need to be undertaken prior to this).

Technical requirements

Components required: Senzing and Elastic Search.

Below are example technical specifications for the Senzing deployment for guidance purposes only. A full technical evaluation will be conducted prior to service delivery. This will be priced on a time-and-materials basis from Our SFIA rate card.

- Single Node Deployments

- Up to 10 million records

- Requirements: 8 cores, 48GB of RAM, 200GB of direct-attached SSD or NVMe storage
- Example AWS instance: i3.2xlarge ~\$0.63/hr
- Approximate ingestion rate of typical data at 100 records per second, 10 million would load in about 1 day.

>>>>>>>>

- Up to 50 million records

- Requirements: 16 cores, 96GB of RAM, 1TB of direct-attached SSD or NVMe storage
- Example AWS instance: i3.4xlarge ~\$1.25/hr
- Approximate ingestion rate of typical data at 200 records per second, 50 million would load in under 3 days.

>>>>>>>>>>>>

- Up to 100 million records

- Requirements: 32 cores, 192GB of RAM, 2TB of direct-attached SSD or NVMe storage
- Example AWS instance: i3.8xlarge ~\$2.50/hr
- Approximate ingestion rate of typical data at 400 records per second, 100 million would load in under a week.

- Multi-Node Deployments

- Multi-node deployments easily support billions of input records. Technical specifications for this volume of records are available on request.

>>>>>>>>>>>>

- Storage Guidelines

- The general guideline on storage planning is to allocate 20kB of flash-based storage per input record. This equates to approximately 1TB of storage per 50M records.

- 20kB is the baseline estimate, but it can increase if your data is very feature-rich and varied. We suggest that you base your storage requirements on an analysis after loading a large sample data source.
- You will also need to account for additional system software, logs, source record files, etc. This can be placed on general-purpose storage.

- Scaling Systems

>>>>>>>>>>>

- Cloud Environment Considerations

>>>>>>>>>>

- Real time operations considerations:

>>>>>>>>>>>>>

- Performance Considerations

Offboarding: data removal

All data is stored in non-proprietary formats in an Elastic Search index. This data will be stored in the Client's Cloud and will be available at all times. At point of contract termination or for any other reason (e.g. specific policy driven schedules), data will be removed from all stores and from the Senzing engine itself. Removal costs will be born by the Buyer and will be as per the data engineering day rate (see SFIA rate card.)

Security

All data is secured as per Buyer's cloud in conformity with any associated security policies. We can provide resources with SC and DV clearances.

Pricing

Please see the separate 'pricing document' and 'SFIA rate card document' for details about software costs and consultancy day rates.