**AMAZON DATA ENGINEER**

**ROLE GUIDELINE**



**Guideline Last Updated**: November, 2017

**Contains Expectations for**: Individual Contributors only

X

**Guideline Notes:** Questions email [de-level-clarification@amazon.com](mailto:de-level-clarification@amazon.com)

# Amazon Data Engineer (DE) Role Description

This guideline contains the general expectations of the Data Engineer role. It describes the most common responsibilities, however given the wide variety of businesses and technologies at Amazon, it cannot capture all expectations. It’s also conceivable that some responsibilities do not pertain to some employees. No two Amazon teams are alike and each is encouraged to develop their own approach to delighting their customers. These realities alter the way Data Engineer is expected to operate and what constitutes success.

Each section has a specific purpose:

* **Section 1: “DE Level Matrix”** is a high-level view of how DE functional abilities map to each level. It is for quick comparisons of level vs. role expectations in hiring debriefs, performance, and promotion discussions.
* **Sections 2-5:** “**What you do**” contains the most detail about DE job level expectations to guide team hiring strategies and performance discussions. Contains a graphic to illustrate the scope and impact for the level.
* **Sections 2-5 “Moving to…”** bullets isolate key skills at the next level that the company would like to see DEs consistently demonstrate before promoting. While it is tempting to use these as a checklist, keep in mind that they are aspects of performance that trigger *consideration* for promotion. Every promotion case is unique; the results you deliver (*and how they are delivered)* also play a role in promotion evaluations.

Your level is acknowledgement that you demonstrate most or all of the skills outlined in both *current* and any *previous* *levels*. Which is to say each level inherits the expectations of previous levels.

## DE Level Matrix

The purpose of the DE Level Matrix is to provide a high-level view of how DE functional areas change by level. The Level Matrix does not include *Amazon Leadership Principles*, as they do not change by level or by role. For those expectations, see <https://w.amazon.com/index.php/LeadershipPrinciples>.

The DE Level Matrix is not to be used in isolation. It is a reference for hiring debriefs, OLR/Promo discussions, and for any other reason where role expectations are discussed in contrast with an individual’s abilities. For hiring, coaching, and performance assessments please use the more detailed guidance in **Sections 3-5**.

### 1.1 DE Functional Areas

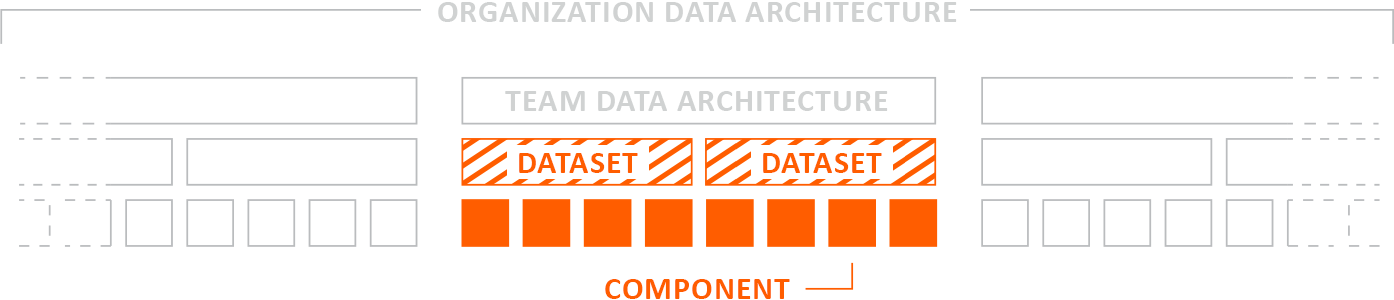
* **Ambiguity**: Level of independence. Degree to which data needs and infrastructure requirements are defined.
* **Scope of Influence**: Influence over data design, architecture, technical strategy, product features, etc. Degree of influence developing others and driving engineering best practices.
* **Advises:** Who DEs collaborate with or influence
* **Technical**: Innovation scope. Degree of difficulty, efficiency, extensibility, data optimization to expect. Ability to deconstruct and simplify problems.
* **Impact**: System size and importance. Impact on architecture. Business Impact.
* **Process Improvement**: Improves data access, data pipeline, and engineering processes (i.e. build, test, release/deploy, monitor, etc.). Focus is on data solution resilience, maintainability, operational excellence, security, and data quality.
* **Knowledge/Judgment**: Understanding of successful design, data structure, integration patterns, and data architecture. Ability to make good decisions that consider the whole picture including balancing difficult short-term and long-term tradeoffs.

### DE Level Matrix

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Dimension** | **L4: DE I** | **L5: DE II** | **L6: DE III** | **L7: Principal DE** |
| **Ambiguity** | Logical data model and requirements are defined. Physical data model is not defined. Will implement and test with guidance. | Logical data model and requirements are not well defined. Data design is not defined. Trusted with more autonomy. Seeks direction at times. | Business problem is defined, **team**[[1]](#footnote-2) technical strategy is not. Logical data model and end-to-end data flow is not defined. Delivers with limited guidance. | Business and/or architectural strategy may not be defined. May not even know what the problem is before starting. Delivers with complete independence. |
| **Scope of Influence** | Self. Collaborates with peers. May mentor DE Interns. | Influence mid-size data solutions/access to dataset(s) in team architecture. Begins to mentor and develop peers. Resolves discordant views. | Influences team data architecture (may extend to related teams) and may influence software design. Actively mentors/develops others. Performs DE technical assessments. | Influences **organization**[[2]](#footnote-3) data architecture. Influences business and technology direction to ensure data integrity. Aligns teams toward simple, coherent designs. Actively develops others. Mentors. Performs Principal DE promo assessments. |
| **Advises** | Peers | Engineers, Product Managers, Manager | Managers, Sr. Managers (Tech and Business) | Director/VP, Broader Tech Community |
| **Technical[[3]](#footnote-4)** | Builds/optimizes physical data model and data pipelines for simple datasets. Troubleshoots existing datasets. Solutions may need refinement. | Builds/optimizes logical data model and data pipelines for difficult datasets in team area. Owns ongoing data quality. Solutions are testable, maintainable, and efficient. | Defines and owns data architecture at the team level Works to simplify, optimize, remove bottlenecks, etc. Solutions are extensible and scalable. | Defines and owns organization data architecture. Solves significantly complex or endemic problems. Solutions are exemplary in terms of robustness, stability, scalability, cost-effectiveness. |
| **Impact** | Impacts data pipeline and dataset quality. | Impacts analytics self-service access to datasets. Increases business effectiveness. | Impacts team data architecture to support analytics use cases. Work yields significant data quality, availability, or business value. | Impacts large-scale data architecture. Uses broad expertise or unique knowledge to refine data pipelines and organization access to data. |
| **Process Improvement** | Measures and improves dataset quality. | Improves self-service access to data, **code**[[4]](#footnote-5) quality, and dependency management. Establishes **SLAs**[[5]](#footnote-6) and defines data certification. Automates manual processes, | Improves business and engineering team process via data architecture, engineering, test, and operational excellence best practices. | Sets the standard for engineering, test, and operational excellence, drives best practices in data architectures/modelling (in organization). |
| **Knowledge (Judgment)** | Data processing language(s).  Learning team data architecture. Actively seeks knowledge and applies to data solutions. | Data design patterns, knows when/when not to use. Understands team data architecture. Makes technical trade-off decisions at dataset(s) level. | Data design best practices. Has detailed knowledge of team data architecture. Makes technical trade-offs between short-term needs and long-term goals. | Core system technologies relevant to organization data architecture. Applies knowledge to invent, evolve, improve, and simplify. |

## 2. DE I

### 2.1 Scope and Impact

****



### 2.2 What you do…

You use technology to solve straightforward problems, seeking input and guidance from team members. Your scope of work is at the small to mid-size component level, building solutions that to enable others answer questions with data. Your work incorporates software engineering best practices, data management fundamentals, and data storage principles. You are able to take well defined requirements, build a solution, and deliver it on schedule.

You are knowledgeable in a variety of strategies for ingesting, modelling, processing, and persisting data. You are able to use one or more query languages (e.g. SQL, HiveQL, SPARQL), schema definition languages (e.g. DDL, SDL, XSD, RDF), and scripting languages (e.g. Perl, Python, KornShell, Scala) to build a data solution. You understand distributed system concepts from a data storage and compute perspective. You understand the differences between data persistence solutions (e.g. HDFS vs RDBMS), data integration techniques (e.g. ETL vs federation), database optimization (e.g. partitioning, distribution, indexing), and join algorithms (e.g. hash vs nested loop). You write secure, stable, testable, maintainable code with minimal defects. You are able to create physical data models and optimize data pipelines for simple datasets.

You participate in team design, scoping, and prioritization discussions. You seek to learn the business context and technologies behind your team’s data architecture. You work effectively with customers (e.g., Business Analysts, Data Scientists, etc.) and other internal partners to identify opportunities/problems. You are familiar with the implications of different implementation decisions (e.g. distribution, normalization). You invent, refine and develop your data solutions to ensure they are meeting the needs of the business and team goals. You are a passionate advocate for your customer.

You assume responsibility for the state of the code you both inherit and produce. You do not put the company at risk (e.g. proper access control, data integrity, auditability, etc.). You get your data models, pipeline designs, and code reviewed. You test your code thoroughly. You classify, store, and handle data in accordance with Amazon policies. You track security risks, mitigate, and/or escalate in a timely manner. You understand the maintenance characteristics, runtime properties, and dependencies of your team’s data solutions and underlying system infrastructure. You clearly document your solutions to ensure ease of use and maintainability by others. You partner effectively with peer SDEs and other engineers. You are able to troubleshoot, research root cause, and thoroughly resolve defects. You handle problems (even when outside your own domain), propose solutions, and either take ownership for their resolution or ensure a clear hand-off to the right owner. You participate in the interview process and help your team train and mentor interns.

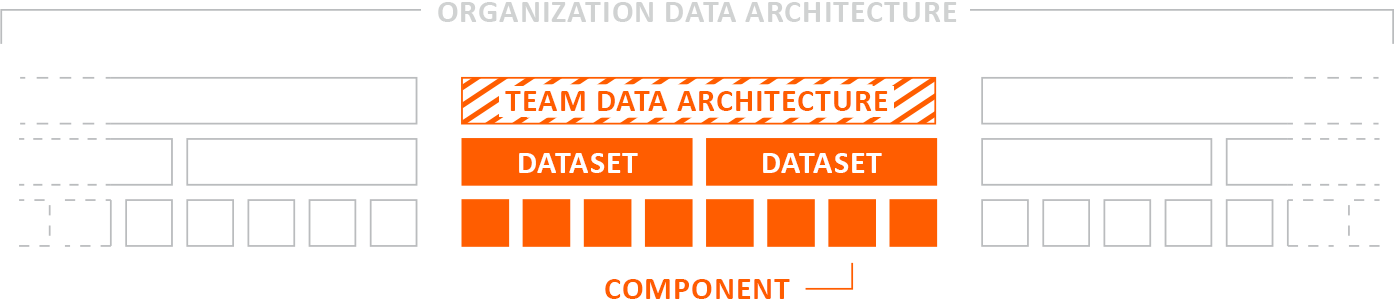
### 2.3 Moving to DE II…

You will be considered for promotion to DE II if you consistently demonstrate a combination of the following:

* You design, implement, and deploy data solutions. You solve difficult problems generating positive feedback.
* You have a solid understanding of data design approaches (and how to best use them).
* You are able to work independently and with your team to deliver data solutions successfully.
* Your work is consistently of a high quality (e.g., secure, testable, maintainable, low-defects, efficient, well documented etc.) and incorporates best practices. Your team trusts your work.
* Your data model and code reviews for tend to be rapid and uneventful. You provide useful review comments for changes submitted by others.
* You focus on operational excellence, constructively identifying problems and proposing solutions, taking on projects that improve your team data architecture, making it better and easier to maintain.
* You make improvements to your team’s development processes.
* You have established good working relationships with teammates and peers. You recognize discordant views and take part in constructive dialogue to resolve them.
* You are able to confidently train new teammates about your customers, your team’s data architecture, how it is constructed, tested, operates, and how it fits into the bigger picture.

## 3. DE II

### 3.1 Scope and Impact





### 3.2 What you do…

You are a significant and autonomous contributor. Your work is consistently of high quality. You solve difficult problems, applying appropriate technologies and best practices. Your focus is at the team level on a major portion of existing or new data architecture (e.g., large or significant dataset, mid-size data solutions). You work with SDEs/other DEs to invent, design and build data solutions that are stable and performant. You create coherent Logical Data Models that drive physical design. Your responsibilities may range from optimizing operational data storage to processing semi-structured data streams to building self-service business intelligence infrastructure for analysts. Whether you specialize in one functional area or work across all of them, your end product is always usable datasets that provides business value.

You are up to speed on recent advances in distributed systems (e.g. MapReduce, MPP architectures, and NoSQL databases). You write code that a DE or SDE unfamiliar with the system can understand. Your solutions are pragmatic. You are proficient in a broad range of data design approaches and know when it is appropriate to use them (and when it is not). You do things with the proper level of complexity the first time (or at least minimize incidental complexity). You consider the legacy of the code you produce, limiting the use of short-term workarounds. You create flexible data solutions without over-engineering. You make appropriate trade-offs, re-use where possible, and are judicious about introducing dependencies. You are efficient with resource usage (e.g., system hardware, data storage, query optimization, AWS infrastructure etc.) You may build tools for efficiently managing data, such as tracking data lineage, ensuring data quality, and improving discoverability of data.

You work on project ideas with customers (e.g., analysts, scientists), stakeholders, and engineer peers. You help balance customer requirements with team requirements. You help your team evolve by actively participating in the code review process, design discussions, team planning, and ticket/metric/COE reviews. You focus on operational excellence, constructively identifying problems and proposing solutions. You take on projects and make enhancements that improve data processes (e.g., data auditing solutions, management of manually maintained tables, automating, ad-hoc or manual operation steps). You work to resolve the root cause of complex problems, leaving data solutions better and easier to maintain than when you found them. You are able to train new peers about how team data solutions are constructed, how they operate, how secure they are, and how they fit into the bigger picture. You foster a constructive dialogue and seek resolutions in a professional way. You help recruit and interview for your team. You mentor and help to develop others.

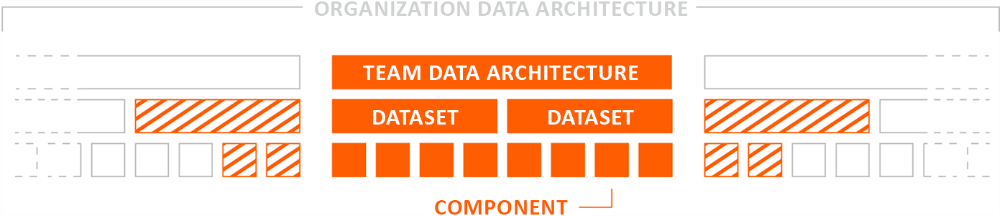
### 3.3 Moving to DE III…

You will be considered for promotion to DE III if you consistently demonstrate a combination of the following:

* You lead the design, implementation, and successful delivery of large-scale, critical, or difficult data solutions involving a significant amount of work. These efforts can be either a new data solution or a refactor of an existing solution. You heavily influence the design and write a significant portion of the “critical-path” code.
* You think in terms of architecture, not just code. You proactively work to improve data quality and consistency. Where needed, you integrate your team’s data solutions with those owned by other teams.
* You influence your team’s technical and business strategy by making insightful contributions to team priorities and overall data approach. You take the lead in identifying and solving ambiguous problems, architecture deficiencies, or areas where your team bottlenecks the innovations of other teams. You make data solutions simpler.
* You are able to communicate your ideas effectively to achieve the right outcome for your team and customer. You harmonize discordant views and lead the resolution of contentious issues (build consensus).
* You lead design reviews for your team and actively participate in design reviews of related software.
* Your code, designs and implementation decisions set a great example to others. You provide insightful code reviews and take ownership of outcome. (You ship it, you own it.) You work very efficiently and routinely deliver the right things.
* You demonstrate technical influence over 1-2 teams, either via a collaborative development effort or by increasing their productivity and effectiveness by driving data engineering best practices (e.g. Code Quality, Data Quality, Logical and Physical Data Modelling, Operational Excellence, Security, etc.).
* You actively participate in the hiring process as well as mentor others - improving their skills, their knowledge of your software, and their ability to get things done.

## 4. DE III

### 4.1 Scope and Impact





### 4.2 What you do…

You are considered a technical leader on your team. You work efficiently and routinely deliver the right things with limited guidance. Your work focuses on ambiguous problem areas in existing or new data initiatives. You take a long term view of your team’s data solutions and how they fit into the team’s architecture. You anticipate data access patterns and remove bottlenecks from their systems. You ensure your team’s data is auditable, available, and accessible. You proactively fix data architecture deficiencies and/or propose larger projects, which may require the work of other teams. You split that work into parallel tasks that can be performed by you and others and then reassembled successfully.

You understand the business impact of your systems and show good judgment when making technical trade-offs between your team’s short-term technology needs and long-term business needs. You are a key influencer in team strategy. You drive mindful discussions with customers and peers. You bring perspective and provide context for current technology choices and guide future technology choices. You understand that not all problems are new (or require new data solutions). You understand the limitations of the systems you work with (e.g., scaling factors, boundary conditions) and the reasons behind existing data solutions *(Q. Why did we build X in this way? What assumptions were made? Do we need to build something else– if so why?).* You make appropriate architectural trade-offs (e.g., *Is it better to build or buy a Business Intelligence technology*? *Do we implement a tiered storage strategy*?). Your code submissions and approach to work are exemplary – your data solutions are easily usable by customers, inventive, secure, easily maintainable, appropriately scalable, and extensible. You build data solutions that are easy for others to contribute to.

You take ownership of team architecture, providing a system-wide view and design guidance. You make things simpler. You drive data engineering best practices (e.g. Data Discovery, Naming Conventions, Operational Excellence, and Data Security) and set standards. You work to resolve the root cause of endemic problems including areas where your efforts unblock innovation of related teams. This may require you to influence decisions made by *other* teams. When confronted with discordant views, you are able to find the best way forward and influence others to follow that path (build consensus). You actively recruit and help others leverage your expertise, by coaching and mentoring in your organization (or at your location). You provide technical assessments for DE II and DE III promotions. You contribute to the professional development of colleagues, improving their technical knowledge and their understanding of data engineering best practices. You ensure your team is stronger because of your presence, but does not require your presence to be successful.

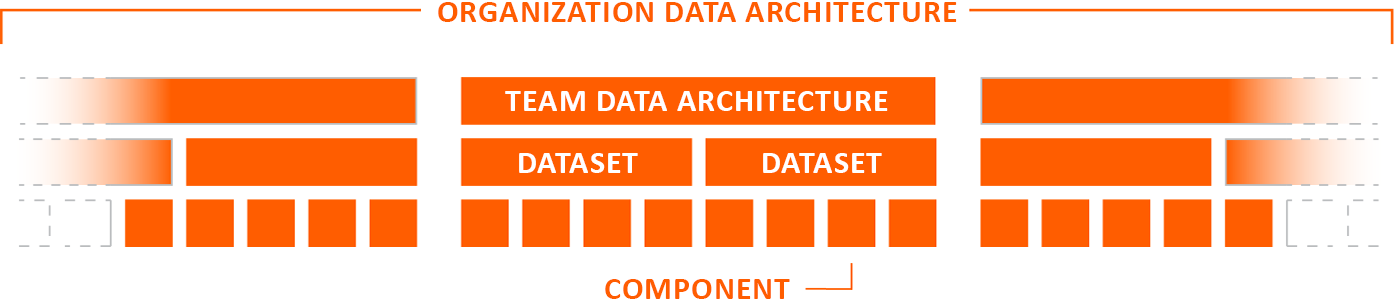
### 4.3 Moving to Principal DE…

You will be considered for promotion to Principal if you consistently demonstrate a combination of the following:

* You take the lead on the design, implementation, and delivery of large-scale or significantly difficult data solutions that have a long-term impact on a product, technology, or architecture.
* Your design approaches and the code you personally contribute is noteworthy in some way (e.g., a significant refactor that simplifies or improves architectural quality, enables us to maintain a competitive advantage, offers significant extensibility, performance, scalability, or has organization-wide operational excellence impact).
* You identify and tackle intrinsically hard problems. (e.g., highly complex, ambiguous, undefined, with less existing structure, or having significant business or security risk or potential for significant impact).
* You deliver artifacts that set the standard in your organization for data engineering excellence, from designs to data models to implementations. Your personal code submissions and reviews of other people’s code are instructive and make the overall code corpus better.
* You lead and actively participate in design reviews, aligning teams across your organization towards coherent architectural strategies. You bring clarity to complexity, probe assumptions, illuminate pitfalls, and foster shared understanding.
* You are a pragmatic problem solver, applying judgment and experience to balance trade-offs between competing interests. You are flexible, adapting your approach to meet the needs of the team, project, and product.
* You actively recruit for Amazon as well as participate in the hiring/interview process.
* You play a significant role in the career development of others, actively mentoring and educating the larger DE community on trends, technologies, and best practices.
* You keep abreast of industry trends. You effectively research and benchmark our technology against other competing software in the industry. You understand that many problems are not essentially new.
* You contribute to intellectual property (e.g., invent, submit patents, etc.)

## 5. Principal DE

### 5.1 Scope and Impact





### 5.2 What you do…

You are a trusted part of the technical leadership of an organization (typically Director level). As a key influencer in planning strategy, you bring business and industry context to technology decisions. You set the standard for data solutions and engineering excellence in your organization. Your data architectures are exemplary in terms of efficiency, stability, extensibility, and the ability to evolve over time. You simplify data access, analyst processes, and technologies. Your data solutions are robust in the presence of failures, scalable, and cost-effective. Your coding practices are exemplary in terms of code organization, clarity, simplicity, error handling, and documentation. You tackle intrinsically hard problems, acquiring expertise as needed. You decompose complex problems into straightforward solutions. You propose projects that may require the work of several *teams* to implement; you divide responsibilities so that each team can work independently and have the system come together into an integrated whole. You are flexible, adapting your approach to meet the needs of the team, project, or product. You solicit differing views and are willing to change your mind as you learn more. You are adept at building consensus.

As a “hands-on” technical leader, you are accountable for portions of the solution architecture, which includes taking ownership of the design and quality of the usability, governance, performance, availability, security and operational aspects of the data solutions built by one or more teams. You split your time between coding, design, and architecture based on where your skills will have the greatest impact (or in response to job requirements). For example, your expertise may be broadly applied – spread across many teams – involved in the technical strategy, design, and delivery of a significant portion of architecture. In this capacity you might look at where we apply the same engineering or analyst effort to get high quality data or focus on closing a gap in a large data architecture. Or you may personally produce code for significant, critical, or demanding software and influence just the few teams close to it. The exact role you play may also change as a larger initiative progresses; during the early phases, broadly influencing several related teams and then subsequently spending weeks or months focusing on a particularly challenging system. Sometimes the mix goes the other way and you spend most of your time broadly influencing multiple teams while occasionally taking a deep dive into a critical or complex area in a particular team.

You amplify your impact by leading design reviews for complex software and/or critical features both in your organization or at your location. You probe assumptions, illuminate pitfalls, and foster shared understanding. You align teams toward coherent architectural strategies. You educate, keeping the data engineering community up-to-date on advanced technical issues, technologies, and trends. You participate, sharing knowledge, and collaborating with other Senior Data Engineers, specifically attending and/or presenting at internal conferences, Principal Engineer community events, and making yourself available to the global tech community. You help managers guide the career growth of their team members by mentoring, coaching, acting as a Principal Advisor, and participating in performance discussions.

### 5.3 Moving to Senior Principal - Tech

Please refer to the **Senior Principal – Tech** role description.

# DE ROLE CLARIFICATION FAQ

# Why are we revising the DE Role Guideline?

In the past few years, industry expectations for the DE job function have shifted. The changes are the result of rapidly evolving technologies and the requirement for robust architectures to pipeline “Big Data”. In a sense, the bar was raised on the DE role. It is now considered a specialized function that builds the infrastructure to answer questions with data using software engineering best practices, data management fundamentals, data storage principles, and recent advances in distributed systems. DE responsibilities now range from optimizing operational data storage to processing complex semi-structured data streams to building self-service business intelligence infrastructure for analysts. A DE can specialize in any one of these functional areas or work across all of them, but their end product is always usable datasets that provide business value.

Amazon role guidance (published in 2009, minimally revised in 2014) did not contain these expectations. It was also ambiguous around the DE scope and technical skill requirements at each level. This ambiguity coupled with out of date expectations has created a significant disconnect with industry DE capabilities. The result was a job internally that contained several types of individuals: those whose main function is to provide reporting or analysis and builders of data models and data architectures. This wide disparity of skill sets created challenges hiring and developing DEs.

Given Amazon has several other job codes for analysis and reporting functions, it was decided to redefine DE to meet industry standards and recommend that employees performing those functions be moved into job codes that reflect their work better.

# What changed with the revised DE role description?

This is a major revision. DE role expectations have been revised to meet industry standards and reflect the skills needed to work with the emergence of Big Data technologies, deprecation of Oracle databases (Rolling Stone, Oracle DW Depreciation), and the impacts of emerging technologies on company data.

Key changes:

* **Scope:** Previously, the primary requirement of a Data Engineer was “*Adopts best practices in data warehousing and business intelligence*”. Business Intelligence is a broad term ranging from reporting infrastructure to reporting visualizations and diagnostic analytics. The DE work group found this wording to be a root cause of the ambiguity and disparate skill sets of our current DE population. **The revised primary requirement meets industry expectations:** *“Accountable and responsible for designing, building, and maintaining data infrastructure(s)”.*
* **Technical Bar:** DE technical expectations now align to the SDE technical bar at each level. This change was needed to meet industry expectations for the role.
* **Exclusions:** The skills required to build self-service reports and dashboards are different from designing, building, and maintaining an efficient, extensible, and scalable data infrastructure. With this revision, DEs will no longer have a primary expectation to create reports or perform data pulls for analytics. This responsibility will now be solely owned by Business Analysts, BI Engineers, IT App Analysts, and Programmer/Analysts.
* **Guideline Format:** Now consistent with Amazon new role guidelines.

The updated DE role description now more clearly defines expectations at each level and the criteria used to consider promotion to the next level. It is applicable to all DEs, regardless of location, organization, or technology focus.

# How did you determine current DE expectations by level?

The DE work group referred to recently revised SDE and DBE role descriptions for level guidance. For expectations we transferred all relevant technical and functional requirements from the old DE role guideline, inspected LinkedIn job requirements, and mined the breadth and depth of Amazon (and industry) experience of DEs and DE leaders involved in the guideline revision vetting process.

# Have expectations changed for any DE level with the revision?

Yes. Level 4-7 DE technical skill requirements now align more closely to industry technical requirements, due to the evolution in technologies which now require a stronger software background and expertise in data modelling, data storage and management fundamentals, data pipeline design, and data architectures.

In addition, previous DE guidance at level 4 (L4) expected *proficiency* in data modelling which is not taught in most universities. Even though SDEs and DEs come from a similar educational background, this created a disparity between L4 DEs who have a median industry experience of 4 years versus 2 years for SDEs (self-reported via the 2017 Tech Survey). To resolve, the L4 requirement is now a basic skill in physical data modelling, which is in a better alignment with how DEs grow this skill. It also allows Amazon to have DE interns and campus recruits which will improve the DE hiring pipeline.

# Is “coding” a skill required for the Data Engineer role?

Yes. DEs need to be able to read, write, and debug data processing and orchestration code (e.g. transformations, scheduling, dependency management, data validation) in order to understand, implement, and maintain datasets. This code is held to the same standards as all the other code written at Amazon (e.g. version controlled, code reviewed, etc.).

# What is the Technical Bar for a DE

The following expectations were pulled from the “What you do…” sections in this guideline. They are isolated here to make it easier for interviewers who are expected to test DE technical criteria. When looking at a level, be sure to include all technical expectations, up to and including that level.

**For Bar Raisers and Hiring Managers, here are specific DE technical expectations by level:**

**DE I:**

* Has engineering background (e.g., CS, CE, or EE degree or equivalent experience).
* Knowledgeable in a variety of strategies for ingesting, modelling, processing, and persisting data*.*
* DE I proficiency in a scripting language (e.g. Perl, Python, KornShell, Scala).
* Is able to use one or more querying languages (e.g. SQL, HiveQL, SPARQL)
* Understands one or more schema definition languages (e.g. DDL, SDL, XSD, RDF)
* Can articulate differences between datatypes (e.g., JSON/NoSQL, relational)
* Is able to create a physical data model
* Is able to optimize data pipes for simple datasets
* Able to dive deeply into technical details (e.g., key dependencies, design choices, operability, etc.) and drive a constructive technical discussion.
* Familiar with the implications of different implementation decisions (e.g. distribution, normalization).
* Understands distributed systems concepts, primarily as it pertains to data storage and compute..

**DE II:**

* Track record delivering data solutions that are correct, stable, and performant. Examples may include optimization of an operational data store, processes for semi-structured data streams, a self-service business intelligence infrastructure, and tools to manage data (e.g., tracks data lineage, ensures data quality, or improves data discoverability). Whatever the specialty, the DE has delivered usable datasets that provide business value.
* Software code and data solutions are high quality. Writes code that a DE or SDE unfamiliar with the logic can understand.
* Proficient in a broad range of data design approaches. Knows when it is appropriate to use them (and when it is not). Does things with the proper level of complexity the first time (or at least minimizes incidental complexity). Can create flexible data solutions without over-engineering.
* Up to speed on recent advances in distributed systems (e.g. MapReduce, MPP architectures, and NoSQL databases).
* Can create coherent Logical Data Models that drive physical design.
* Understands how to make appropriate data trade-offs. Can balance customer requirements with technology requirements. Knows when to re-use code. Is judicious about introducing dependencies.
* Understands how to be efficient with resource usage (e.g., system hardware, data storage, query optimization, AWS infrastructure etc.)
* Knowledge of engineering and operational excellence best practices. Can make enhancements that improve data processes (e.g., data auditing solutions, management of manually maintained tables, automating, ad-hoc or manual operation steps).

**DE III:**

* Track record delivering large data solutions (e.g., data model, data architecture, data flow design, tool) in difficult or ambiguous data areas (new or existing systems).
* Code submissions and approach to work are exemplary. Data solutions are easily usable by customers – inventive, secure, maintainable, scalable, and extensible.
* Builds data solutions that are easy for others to contribute to. Knows how to make data auditable, available, and accessible.
* Has significant knowledge/experience with data design approaches and industry technologies.
* Is able to evaluate end-to-end data designs for strengths and weaknesses (data quality, scalability, latency, security, performance, data integrity, etc.).
* Can anticipate data access patterns and remove bottlenecks. Asks the right questions and drives the right technical solution(s).
* Is able to split development work into parallel tasks that can be performed by them and others and reassembled successfully.
* Is able to influence team technical strategy. Understands that not all problems are new (or require new data solutions). Is able to make appropriate architectural trade-offs (e.g. Build or buy a Business Intelligence technology? Tiered storage strategy?). Shows good judgment when making technical trade-offs between short-term technology needs and long-term business needs
* Can take ownership of a team’s data architecture and make it simpler. Has proactively fixed an architecture deficiency. Resolves root cause.
* Is able to drive data engineering best practices (e.g. Data Discovery, Naming Conventions, Operational Excellence, and Data Security) and set standards.
* Understands system limitations, scaling factors, boundary conditions, and/or the reasons for architectural decisions *(Q. Why did we build X in this way? What assumptions were made? Do we need to build something else– if so why?).*

**Principal Data Engineer**

* Hands-on technical leader. Sets the standard for data engineering excellence in an organization.
* Data architectures are exemplary in terms of efficiency, stability, extensibility, and the ability to evolve over time.
* Data solutions are robust in the presence of failures, scalable, and cost-effective.
* Coding practices are exemplary in terms of code organization, clarity, simplicity, error handling, and documentation.
* Is able to tackle intrinsically hard data problems, acquiring expertise as needed.
* Can propose projects that may require the work of several development *teams* to implement. Is able to divide responsibilities so that each team can work independently and have the system come together into an integrated whole.
* Is able to take ownership of the design and quality of the usability, governance, performance, availability, security and operational aspects of the data solutions built by one or more teams.
* Can lead design reviews for complex or ambiguous problems requiring data solutions. Is able to probe assumptions, illuminate pitfalls, and align teams toward coherent architectural strategies.
* Up-to-date on advanced technical issues, data technologies, and trends.
* Has deep knowledge of system technologies in data domain and/or broad understanding of company systems/technologies. Applies this technical knowledge to invent, evolve, improve, simplify, etc.
* Able to identify gaps/opportunities in or between data architectures in systems or regions (e.g., services, workflows, tooling) to satisfy the needs of data customers
* Decomposes complex processes into straight-forward solutions, often inventing new ones.
* Reduces coupling between teams. Looks at where we apply the same engineering effort year-over-year to determine what requires us to exert that effort multiple times. Drives architecture or organization changes to enable teams to work independently and/or achieve a significant efficiency improvement.

# In what way is a coding interview for a DE different than for an SDE?

SDE coding interviews focus on object-oriented imperative programming languages (e.g. C++, Java) while DE coding interviews focus on declarative programming languages (e.g. SQL, SPARQL) and functional ones (e.g. Scala, Python, Perl). DEs are expected to understand how to optimize the distribution, partitioning, and massive parallel processing of high-level data structures (e.g. tables, HDFS), but are not expected to understand dynamic memory allocation of programs and low-level data structures (e.g. arrays, hash maps, trees) like SDEs do. Both interviews should cover code quality (e.g. naming conventions, unit testing, maintainability, efficiency), software design best practices (e.g. extensibility, availability, scalability, abstraction), and operational excellence. DE coding interview questions must give the candidate the opportunity to demonstrate the ability to process a variety of data types, orchestrate across disparate sources, and deliver it so it is trustworthy and easy to use. At higher DE levels, interviews are very similar to SDE however, the scope of the design and architecture questions would be on data solutions.

# How is the DE role different than the SDE role?

While SDEs are expected to understand the basics of the design, storage, and maintenance of the data their software produces, DEs specialize in those areas. At a high level, DEs build the infrastructure to answer questions with data, using software engineering best practices, data management fundamentals, data storage principles, and recent advances in distributed systems (e.g. MapReduce, MPP architectures, NoSQL databases). DE responsibilities can range from optimizing operational data storage to processing semi-structured data streams to building a self-service business intelligence infrastructure for analysts to use. They may build tools for efficiently managing data, such as tracking data lineage, ensuring data quality, and improving discoverability of data. However they are applied, their end product is always usable datasets that provides business value. They ensure data is auditable, available, and accessible. For example, a DE may help specify and tune the persistence layer that is being built for a product or service and may facilitate collaboration between SDEs and Product Managers to ensure all required data is securely persisted and will be available for downstream system usage.

# How is the DE role different than other data roles?

Amazon has several different job functions related to data. At a high level, these are the use cases and the roles they map to:

|  |  |  |  |
| --- | --- | --- | --- |
| AREA | RESPONSIBILITIES | Pre-revision Role | Post-revision Role |
| Database and Persistence Solutions | Designs, builds, and delivers original database solutions, appliances, products, or tools (e.g., Redshift, S3, NoSQL Databases such as DynamoDB) | * DBE | * DBE |
| Database Management | Administers databases. Has deep expertise in maintaining and scaling database products (e.g. Redshift, MySQL, Oracle). Creates database scripts, indexes, partitions, shards, monitors, backups, logs, and metrics. Responsible for database availability, integrity, and compliance (SOX). Performs database recovery. | * DBA | * DBA |
| Data Infrastructure | Builds and maintains the infrastructure to answer questions with data, using software engineering best practices, data management fundamentals, data storage principles, recent advances in distributed systems, and operational excellence best practices. Builds datasets that analysts and scientists use to generate actionable insights. | * DE * SDE * BIE | * DE * SDE |
| Business Reporting | Builds and manages automated reporting framework(s) including WBR, MBR, flash reports, dashboards for business and other functions (e.g., Operational Metrics, etc.). Supports day-to-day decision making necessitated by the business and its various cycles. Expertise in Excel and SQL. Level of coding ability or required expertise with third-party technologies and external visualization tools (Tableau, Microstrategy, Quikview, Quicksight) determines job code placement. | * BA * BIE * DE | * BA * BIE * It App Engineer * IT App Analyst * Programmer /Analyst |
| Data Science | Performs statistical analysis and develops algorithms. Provides predictive data models to support business decisions. Degree and expertise determines job code placement. Works with a variety of math/stats methods (e.g. Matlab, R, Stata). May work with scripting languages (e.g. Python). Experience with “Big Data” tools (e.g., Hadoop, Spark, MongoDB, and other Map Reduce) not required, but should know when to use such tools. Excel and SQL can be learned on the job. Delivers insights via whitepapers and visualization tools. May design interactive dashboards (e.g., in Tableau) to reinforce analytical findings, but this is not a dominant portion of their work. Does not build datamarts. | * BIE * DE * Research Scientist * Economist | * Data Scientist * Research Scientist * Economist |

# DE and DBE seem similar, how exactly are they different?

DBEs are storage specialists that work with service and product development teams to build database and other data persistence solutions. DEs produce and maintain datasets leveraging these products along with other persistence and compute technologies (e.g. HDFS, EMR). DBEs operate at the persistence level while DEs work in the data processing layer.

To address the potential for confusion between these two roles, a role comparison is below.

**High Level**

|  |  |  |
| --- | --- | --- |
| DATA ENGINEER |  | DATABASE ENGINEER |
| Responsible for designing, developing, troubleshooting, evaluating, deploying, and documenting data management and business intelligence systems for customer and internal applications. | **Core Responsibility** | Responsible for designing, building, and delivering database solutions, appliances, products, or tools. |
| In-depth understanding of one or more engineering subject areas as it relates to data storage and movement solutions: data warehousing, enterprise system data architecture, data design (e.g., Logical and Physical Modeling), data persistence technologies, data processing, data management, and data analysis. | **General Knowledge** | In-depth understanding of one or more engineering subject areas as it relates to data persistence solutions: database systems theory, distributed systems, database design, database technologies, operating systems, hardware, and system architecture. |
| Understands the software development lifecycle. Ensures software development best practices related to persistent data storage are observed[[6]](#footnote-7). Builds secure, available, scalable, stable, and cost-effective data solutions using data storage technologies, distributed file system, data processing, and business intelligence best practices. | **Applied Knowledge** | Understand the software development lifecycle and engages with team to ensure software development best practices related to persistent data stores are observed[[7]](#footnote-8). Build stable, scalable, secure, available, and cost-effective database solutions out of database, hardware, OS, storage, and networking technologies. |
| Data storage and processing subject matter expert (SME) and key member of software team(s) that provide data management and data warehousing solutions (e.g., creating new or adapting features to enhance existing DW/ETL or other technologies, designing new ETL products). Helps engineering teams understand their data and processing options to build effective reporting solutions. | **Engineering Function** | Data persistence subject matter expert (SME) and key member of software teams implementing data storage solutions (e.g., adapting existing technologies, adding features to enhance existing database management software, or designing new database products). Help engineering teams understand their data and processing to engineer data storage solutions that are appropriate to the situation. |
| Typically holds a Bachelor’s Degree in Computer Science or related field. | **Education** | Typically holds a Bachelor’s Degree in Computer Science or related field. |

**Key Responsibility Detail**

|  |  |  |  |
| --- | --- | --- | --- |
| RESPONSIBILITY | DATA ENGINEER | OVERLAP | DATABASE ENGINEER |
| System Architecture | * SME for relational and non-relational data management systems. * Expert knowledge in data design that meets the availability, scalability, and capacity demands of the business | * Reviews system designs for performance, security, scalability, and high availability * Expert knowledge of distributed systems and high availability database architectures | * SME for relational and non-relational database management systems. * Influences the design of future database solutions and technologies * Standardizes design patterns and usage of persistence solutions that can be adopted across multiple use cases and customers |
| Requirements Gathering | * Gathers and implements business and technical requirements for changes to data models, ETL pipelines, and decision support systems. | * Excellent communication skills * Writes or contributes engineering specifications for projects based on business requirements | * Gathers and implements business and technical requirements for changes to software applications or database management software |
| Logical Design, Physical Design, DDL | * Creates new logical models and converts into physical models * Creates new physical schemas and modifies existing schemas based on business requirements * Executes data/schema modifications in complex systems with minimal-to-no downtime or service impact |  | * Designs new patterns for software engineers to access and manage data (e.g. how to store time series data in a given database technology) * Work leads to the creation of new reusable database features. |
| Query Languages | * Develops data models that answer various queries without causing operational or maintenance overhead * Designs self-service business intelligence solutions for analysts and scientists | * SME for complex data access patterns and performance optimization | * Develops solutions to query design patterns (e.g., pagination, optimistic locking) that can be re-used across multiple use cases * Designs query visualization and optimization solutions for database systems |
| Coding, Scripting | * Writes code (e.g., SQL, Python and modern programming languages for data analysis like SCALA and R) * Proficient at automating database creations, migrations, and upgrades | * Performs code reviews * Refactors code for simplicity and extensibility | * Knowledge of modern programming languages * Understands how software development languages interact with database management systems via connection pooling, mapping technologies, etc… |
| Deployments, Migrations, Changes | * Designs solutions and tooling to execute automated data workflows and ETL pipelines |  | * Designs solutions and tooling to execute automated database deployments, upgrades and migrations |
| Operations and Issue Resolution | * Responsible for the availability, performance, and correctness of datasets in reporting dashboards. * Evaluates and drives changes to existing data models for scaling big data related problems. |  | * Responsible for the stability, scalability, security, and availability of database systems built by their team. * May be responsible for problem resolution that impacts software teams using their database technologies. * Evaluates and drives changes to database and software architectures to address recurring issues or limitations |
| Technical | * In-depth understanding of data storage, data processing, data management, and data analysis technologies * Able to develop proof-of-concept prototypes using data technologies * Able to integrate multiple datasets or sources into a leverageable dataset | * Basic knowledge of database administration, systems architecture, and software development * In-depth knowledge of database design and database technologies * Able to build/support databases using existing database management software * Ability to optimize database systems for performance, security, and reliability | * In-depth understanding of data architecture * Able to coach software development teams on how to efficiently store, access, and modify data |

1. **Team** is defined as the group reporting to a line manager [↑](#footnote-ref-2)
2. **Organization** is defined as the groups reporting to a Director (or a VP in smaller organizations) [↑](#footnote-ref-3)
3. **Hiring Managers/Bar Raisers** – **See FAQ #6 for the technical requirements of each DE level** (page 9 of this guideline). [↑](#footnote-ref-4)
4. **Code** is defined as a series of program instructions that executes or automates computer commands. In this document it refers to data processing and orchestration commands (e.g. transformations, scheduling, dependency management, and data validation). [↑](#footnote-ref-5)
5. **Service Level Agreement** (SLA) [↑](#footnote-ref-6)
6. A DE does write software code, however their coding bar is somewhat different from an SDE. [↑](#footnote-ref-7)
7. A DBE does not need to meet the SDE coding bar [↑](#footnote-ref-8)