Chapter 17

Process

17.1 WHAT IS WORKFLOW MANAGEMENT?

Many of the things you do can be considered as workflows. Think about the creation of a work product such as this book. Writing the book involves a sequence of activities. I write each chapter in turn (not necessarily in reading order) then review and modify those chapters as new ideas occur to me. When the book's text is complete, it is copy edited. This can involve iterations of review and discussion with the copy editor. Eventually the book is printed so it can be put on sale. The work product I am producing goes through various states such as "in planning," "being written," "under review," "published," and "on sale." The work product can also jump back into previous states. For example, reviewers' comments could put the book back into a state of "being written" as their comments are incorporated into the book.

A workflow then is a set of defined activities on a work product that take place in a prescribed order. Each activity moves the work product into different allowed states such as the "being written" state in our book writing example.

A workflow management *system* is a process or a software tool for monitoring activities and states of work products and controlling the allowed transitions between those states.

17.2 WORKFLOWS IN ANALYTICS

The roadmap through much of this book is based on the Guerrilla Analytics workflow of Figure 59. This is the sequence of activities that is applied to data from extraction through to delivery as team work products and reports. As you are well aware by now, there are various activities that the analytics team member engages in. We have seen activities such as loading data, coding, testing, and presenting results. We could map out all the nuances and checks of data loading and data testing. We could specify the steps to take when choosing and validating a model. Any of these activities could be described in a workflow and therefore could be managed with a workflow management system. To specify every possible analytics activity with a workflow would be overwhelmingly complex and would hamper the agility required for Guerrilla Analytics. In keeping with the principle of light-weight processes, this chapter focuses on

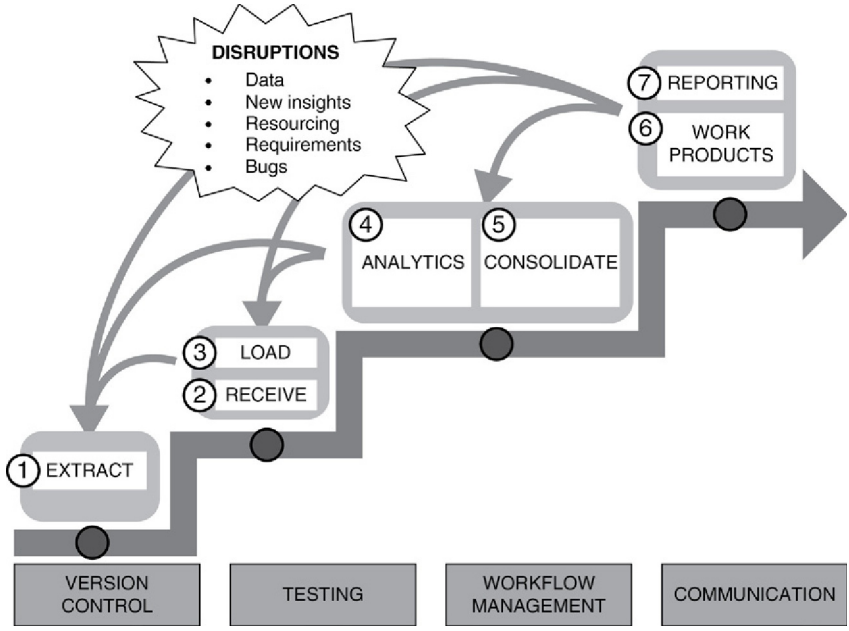

FIGURE 59 The Guerrilla Analytics workflow

the key activities that benefit from a minimal workflow management. As illustrated in the Guerrilla Analytics workflow, workflow management is one of the supporting functions of all of these activities.

17.2.1 Types of Workflow

There are three basic types of workflow management needed in Guerrilla Analytics projects.

- **Data receipt and load:** This is effectively the management of data logging and data tracking. Data gets a special workflow of its own because of its importance. Data loss has very serious repercussions for the team and the customer. As mentioned on many occasions, losing data provenance is the root cause of team inefficiency. The team needs to know where data came from, when it arrived with the team, where it is stored on the file system, and where it was loaded to in the Data Manipulation Environment (DME). Workflow management can help with this.
- **Producing work products:** These are all the ad-hoc analytics activities engaged in by the team. Specifically they are the activities at data extraction, coding, testing, and reporting. As mentioned before, you do not want to delve into the detail of each activity but you do need to know which activities are open, in review, and delivered as well as who is working on what.

- **Programming a build:** These are the build development and testing activities engaged in by the team. The workflow management of build activities is closest to workflow management in a traditional software engineering project (Doar, 2011). The build will be architected into components and these will be coded by team members. Bugs will be reported against versions of a build and these bugs have to be repaired. Releases of builds and versions of new builds need to be coordinated. Workflow management excels here.

The three workflow types described above have been sufficient to coordinate all my teams' activities in very complex and fast-paced projects.

17.2.2 Common Workflow States and Transitions

Although the various analytics workflows differ in the information they need to track, they can all be described by the same high-level activities and states illustrated in Figure 60. We will first look at this common structure before describing how the activities differ in each workflow type.

The workflow begins when a new work product arrives with the team. In data logging, this is a new delivery of data. For producing work products, it is a new piece of work the team has been asked to do. For build activities it will be some feature that needs to be implemented or a bug that needs to be fixed in the build.

The team member who picks up this activity from the customer describes it with appropriate detail and places the work product into a state of Open. The work is now on the team's radar.

Once Open, the work product remains in this Open state until one of two things happen. First, being a very dynamic project, it is quite possible that

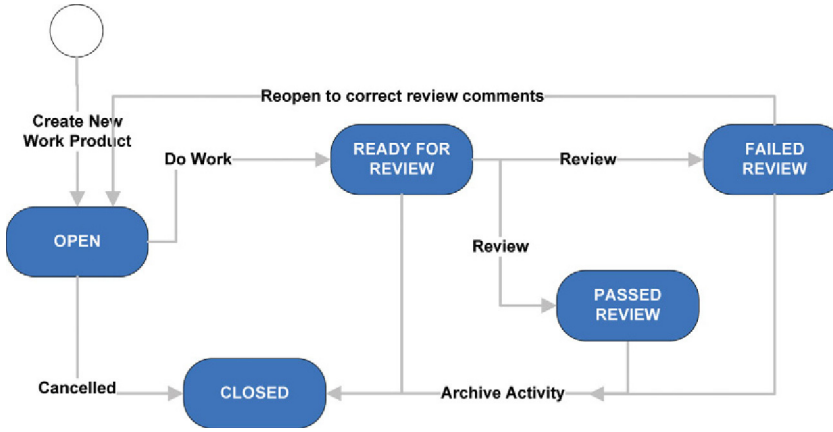

FIGURE 60 A simple workflow

somebody will decide the work product is no longer needed and so the workflow can be closed down. Alternatively, a team member can begin the activity needed to complete the work product. This could be coding, writing, testing, or any of the myriad of team activities. When this is completed, the team member marks the work product as "Ready for Review." Somebody now needs to pick up this work product and review it for correctness.

The review process places the work product into one of two states. Either the work product passes review or it fails review. A failed review leaves the work product in a state of Failed Review. The original team member can then pick up this work product to make changes to it. This pushes the work product back into the original state of Open. The process begins again. If the work product passes review it can go on to be delivered to the customer (internal or external).

What we have here is a minimal set of states and transitions that describe any team activity and review.

17.2.3 Common Information

The analytics workflows are actually very similar and have a lot of descriptive information in common. This descriptive information is needed when the work product is created, when it is worked on, and when it is reviewed. Some typically useful information to track includes the following.

- **Creating the work product:** This is the initial registration of the work product in the workflow tracking system.
  - The customer who requested this work product. This facilitates rapid interaction with the customer and allows themes to emerge.
  - The project work stream that will use the work product.
  - When the work product was requested.
  - The UID of the work product.
  - Which team member is assigned to complete the work product.
- **Completing the work product:** This is the activity of doing the actual work on the work product. Keeping with the Guerrilla Analytics principles, interruptions to doing actual work should be minimized. That said, it is useful to track the following information while completing a work product.
  - Any helpful comments on lessons learned, exceptions, and other context that would help somebody else understand the work product. For example, in a modeling activity, a team member should note their observations about the data and their rationale for choosing a particular model. This is helpful if the original executing team member is no longer on the team or is not available.
  - Any conversations with the customer that helped shape or change the work product as originally described. This is useful if particular design decisions and interpretations of outputs are questioned in the future.
- **Review of the work product:** This is the activity of reviewing the completed work product. Here it is important to capture.

- Who reviewed the work product. This person is different from the work product executor.
- When the reviewer did their review.
- Any comments and feedback on the work product. It is important to capture this information so the evolution of the work product can be understood and also so that junior team members have a record they can refer back to and learn from.

The information above greatly helps in tracking and understanding a work product as well as in any handover of work to other team members.

Given that there are differences between the three types of analytics workflow (data tracking, work products, and build activities), so there are some differences in the information that should be captured about those workflows.

17.2.4 Specific Information for Data Receipt and Load

With Data Receipt and Load, you should track as much of the following key information as possible.

- **Data description:** A short description of the data and any issues reported at delivery time. Some examples include the source system the data came from, the purpose of the data and any known issues or limitations of the data reported by the provider.
- **Delivery details:**
  - Who delivered the data to the team and who in the team received the data? This helps if there are follow-up questions about the data.
  - When did this happen?
  - Any associated communications around the data delivery.
- **Data storage and Load:**
  - Where was the data stored on the file system?
  - Where was the data loaded to in the DME? When loading data files into the DME, you can enforce that the data UID is first in the dataset name. This ensures all raw data list sequentially in the DME and aids in referencing back to the workflow management system.
  - Any issues encountered during data load.
- **Related work:** A link to the work product UID where checksums and other related data tests were performed.

17.2.5 Specific Information for Builds

Build workflows are closest to traditional software development efforts. As such they benefit from incorporating information such as:

- **Bug tracking:** A Bug ID from a bug tracking system when a work product involves fixing a bug.

- **Version control:** A version name and number of the build against which a work product is being applied.
- **Related development:** A link to any associated test scripts and code files.
- **Change control:** Details of the changes requested and features implemented in a particular build version. This aligns expectations as to when new and updated build datasets will be available.

Again, this tracking information is far from the detail offered and used with traditional software engineering projects. Remember, you are not in a traditional software engineering project. You are producing a data build, not a data warehouse and you need to do that in a Guerrilla Analytics environment. What I have listed here is the bare minimum you need to track to get the maximum improvement in your team's resilience to disruptions.

17.3 LEVELS OF REVIEW

A key team activity is to review one another's work. The high level example of workflow from Figure 60 shows only one level of review. In reality there is often a need for two levels of review in analytics projects. This is especially the case for very high profile work products such as reports.

Peer review is where team members review one another's work. Peer review is typically quite technical. It helps identify errors in analytics code or errors in an analytics approach. It also helps more junior team members benefit from the experience and coaching of more experienced team members.

Supervisor review is where an activity is given a second check usually at a higher and less technical level than a peer review. Supervisor reviews focus on management concerns. That is, whether the work product meets customer requirements, whether it complies with the team's processes, and whether data provenance is preserved.

The high-level workflow described by Figure 60 has only one level of review for simplicity. Adding supervisor review would be straightforward. The amount of required review will depend on the importance of the work product.

War Story 16: Team Spirit

Dave was brought into a fairly broken and distressed team on a very high-pressure project. Team members were disconnected from the customer and morale was very low. Being a Guerrilla Analyst, Dave's initial objective was to get team processes in order so he could achieve data provenance and better coordinate the team.

He started by introducing the practice tips and processes you have been reading about throughout this book. However, from the start there was resistance from some team members. As a new manager of the team there was suspicion of his "new rules." To the team they probably seemed whimsical and unnecessary. "Put all my data here and name it this way? Who is this guy?!." Of course they were

not in the project steering meetings where Dave was getting an earful about the lack of coordination and quality in the team and how he had better get to grips with his new role and fix it fast.

Dave began by reviewing work products himself so he could get a handle on the project, the data, and whether the team was doing things in a uniform way as required. In the early days, there was lots of feedback and work products went through several iterations until they met Dave's bar of data provenance. This of course wasn't sustainable with 10 direct reports – he was going to need to delegate these reviews. Also, the team was going to need to understand why they had these rules and processes and follow them because it was the right thing to do.

Dave identified the senior analysts in the team and took them aside individually. He explained how they were role models for the junior team members and how their expertise was important to the success of the project. All expressed a desire to grow and agreed that coaching others and imparting their wisdom was an important part of that growth. Dave then assigned them as first-level reviewers of work products. Anybody creating a work product had to consult with senior team members to find a suitable reviewer who was free. The result was apparent within days. Suddenly, when confronted with a mess of code in many locations, disappearing data, undocumented work products, etc., team members began to understand the importance of the simple rules, conventions, and processes. Reviewers understood how frustrating it was when every one of their reviews began with a long walk-through of a work product and all its unique and ad-hoc nuances. Equally, reviewees would be proud to pass a review from a senior they admired and always strived to get their work product through on the first attempt. Team spirit and data provenance improved and the team was freed up to start adding real value to the project instead of chasing their own tails.

17.4 LINKING WORK PRODUCTS

You have seen how many of the team's work products are associated with one another or follow on from one another in some way. For example, a work product involving checksums of data is obviously associated with the work product that involved logging and loading that data into the DME. Any general work product may be a follow-on or tangent from a previous work product. It is important to be able to capture these linkages as it again helps track and understand the history of a work product as well as identifying customers and team members who were previously associated with a related work product. Since all work products have a UID, linking is simply a matter of listing a work product's linked UIDs and the reason for their linkage. Here are some examples.

- This data test is linked to the follow data import UID.
- This build feature is a fix of the bug with the UID 516 and updates the feature implemented under UID 389.
- This work product is UID 658 and is a report that uses a customer list loaded under UID 380.

17.5 CLASSIFYING WORK PRODUCTS

There are various ways that work products can be classified depending on project needs. I have found the following classifications helpful.

- Be able to distinguish work products by whether they are for internal or external delivery. Customer work products are always more sensitive than ones delivered only to the analytics team.
- Be able to identify work products that have gone into reporting and the particular report they were used in. This is because these work products are particularly sensitive. Reporting and the challenges of data provenance are discussed in detail in a chapter Chapter 10.
- On larger projects, it may be helpful to classify work products by their project work stream.

17.6 GRANULARITY

Probably the biggest decision you will face when putting workflow management into practice is the level of granularity at which you describe the three workflow types. Granularity refers to the number of discrete states and activities that the workflow tracks. The simple analytics workflows put forward in this chapter have the minimal number of states I have found necessary in analytics projects. Those projects have ranged from 3 to 12 team members with local and off-shore resources. These workflows balance the time it takes for the team to update the workflow management system with the desire to coordinate and report on team activities in detail.

Some workflow management software will divide a work product into finer grain states such as "Opened," "Reopened," "In progress," "On hold," "In Review," "Ready for Delivery," etc. with the aim of tracking hours spent in each activity. Think carefully before doing anything more granular than the workflows described in this chapter as maintenance of the work product states in the workflow management software is an administrative overhead on the team. Heavy processes in a Guerrilla Analytics environment are doomed to fail.

17.7 WHEN TO USE WORKFLOW MANAGEMENT

The following considerations should influence your decision to use workflow management in your team and the type of workflow management to implement.

- **Team size:** Once a team has more than about three members, it quickly becomes difficult to have a view of what the team is doing. Which work products are being worked on? Who is doing the work? Is everything being reviewed? Has every customer request been picked up by somebody? What's the next thing I need to do? If a team is going to be of even a moderate size, consider using some basic workflow management – even if that is a simple spreadsheet shared by the team.

- **Project dynamism:** If the project is very dynamic with many work products ongoing then requests for work can get overlooked. A workflow management solution can help capture everything that the team must do, even if they are too overloaded to begin that work immediately.
- **Team experience:** If the team has relatively inexperienced members, their work should be reviewed more thoroughly than other team members. Tracking these reviews and capturing feedback for the benefit of inexperienced team members is best done in a workflow management process.
- **Life of project:** If a project is going to last more than 6 to 8 weeks then two factors come into play. It is more likely that the team composition will change both on the analytics and customer side. This means that the history of past work products could be lost. Twelve weeks later, how many of the team will remember why the build does not have a particular feature that seems obvious and necessary? A workflow management system captures the conversations and project folklore so that the context of past decisions and work products can be retrieved. Second, handovers between team members are greatly facilitated if the leaving team members can simply look up all the work products they produced.
- **Importance of traceability:** In some types of project, it is imperative that every piece of work and the team members associated with it can be tracked. Any work that contributes to legal cases, for example, must be completely traceable. In these types of project, workflow management is very beneficial.
- **Geographically dispersed teams:** Geographically dispersed teams are common in today's globalized world. Off-shoring, near shoring, and customers with operations in a variety of locations can all cause teams to be physically dispersed. A workflow management tool helps these teams coordinate. It works almost like a notice board where teams can communicate and coordinate their tasks with one another even though they might have little overlap in their working hours.

Clearly these considerations will influence the sophistication of workflow management you put in place. Guerrilla Analytics projects have used everything from simple spreadsheets to full-blown web-based applications as project needs dictate.

17.8 WRAP UP

This chapter has discussed workflow management for Guerrilla Analytics. Having read this chapter, you should now understand.

- What is workflow management?
- What are the types of workflow management needed in Guerrilla Analytics?
- A common high-level workflow that describes all Guerrilla Analytics activities.

- The common information that should be tracked in workflow management for Guerrilla Analytics.
- Tracking information that is particular to each of the workflow types.
- Work product review and how this can be implemented with one or more levels of review.
- The concept of linking work products and why this is often useful.
- Some common classifications of work products.
- The factors that influence a decision to use workflow management in a Guerrilla Analytics project.